日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年１２月２７日

出願番号
Application Number: 特願２００２－３７８９５６

[ST.10/C]: [ＪＰ２００２－３７８９５６]

出願人
Applicant(s): 株式会社日立製作所

２００３年　５月１３日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００３－３０３５５９１

【書類名】　特許願
【整理番号】　H02016111A
【あて先】　特許庁長官　殿
【国際特許分類】　G06F 15/16
【発明者】
【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所　システム開発研究所内
【氏名】　細谷　睦
【特許出願人】
【識別番号】　000005108
【氏名又は名称】　株式会社　日立製作所
【代理人】
【識別番号】　100075096
【弁理士】
【氏名又は名称】　作田　康夫
【電話番号】　03-3212-1111
【手数料の表示】
【予納台帳番号】　013088
【納付金額】　21,000円
【提出物件の目録】
【物件名】　明細書　1
【物件名】　図面　1
【物件名】　要約書　1
【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　高可用ディスク制御装置とその障害処理方法及び高可用ディスクサブシステム

【特許請求の範囲】

【請求項１】

ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、記憶装置とのインターフェースを有する複数のディスクインターフェース部と、前記記憶装置に対しリードまたはライトされるデータを一時的に格納する複数のキャッシュメモリ部を有し、前記複数のホストインターフェース部の各々は、前記ホストコンピュータとのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、前記複数のディスクインターフェース部の各々は、前記記憶装置とのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行するディスク制御装置であって、

前記複数のホストインターフェース部と前記キャッシュメモリ部、及び、前記複数のディスクインターフェース部と前記キャッシュメモリ部との間が１つ以上のスイッチで構成されるスイッチ網を介して接続され、前記ホストインターフェース部、前記ディスクインターフェース部及び前記キャッシュメモリ部は、前記スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を有し、前記スイッチは前記スイッチ網内での経路を前記IDを用いて指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、さらに前記複数のキャッシュメモリ部は、該複数のキャッシュメモリ部における障害発生の有無を監視するための障害監視機構と前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を有することを特徴とするディスク制御装置。

【請求項２】

前記パス制御機構は前記複数のキャッシュメモリ部における障害発生時に障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御することを特徴とする請求項１記載のディスク制御装置。

【請求項３】

ホストコンピュータとのインターフェースを有する複数のホストインターフェ

ース部と、記憶装置とのインターフェースを有する複数のディスクインターフェース部と、前記記憶装置に対しリードまたはライトされるデータを一時的に格納する複数のキャッシュメモリ部と、ホストインターフェース部及びディスクインターフェース部と前記キャッシュメモリ部との間のデータ転送に関する制御情報を格納する複数のリソース管理部を有し、前記複数のホストインターフェース部の各々は、前記ホストコンピュータとのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、前記複数のディスクインターフェース部の各々は、前記記憶装置とのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行するディスク制御装置であって、

前記複数のホストインターフェース部、前記複数のディスクインターフェース部が１つ以上のスイッチで構成されるスイッチ網を介して前記キャッシュメモリ部と接続され、前記複数のホストインターフェース部、前記複数のディスクインターフェース部が前記スイッチ網を介して前記リソース管理部に接続され、前記ホストインターフェース部、前記ディスクインターフェース部、前記キャッシュメモリ部及び前記リソース管理部は、前記スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を有し、前記スイッチは前記スイッチ網内での経路を前記IDを用いて指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、さらに前記複数のキャッシュメモリ部と前記複数のリソース管理部は、その障害発生の有無を監視するための障害監視機構と前記スイッチ内の前記フォワーディングテーブルを制御するためのパス制御機構を有することを特徴とするディスク制御装置。

【請求項４】

前記パス制御機構は前記キャッシュメモリ部または前記リソース管理部における障害発生時に障害部位を回避するように前記スイッチ内の前記フォワーディングテーブルを制御することを特徴とする請求項３記載のディスク制御装置。

【請求項５】

前記キャッシュメモリ部における障害を監視するための前記障害監視機構が、前記キャッシュメモリ部の中にかえて、前記リソース管理部の中にあることを特徴とする請求項３または４記載のディスク制御装置。

【請求項６】

前記障害監視機構が、前記ホストインターフェース部及び前記ディスクインターフェース部でのリードまたはライトとともに動作することを特徴とする請求項１乃至５のいずれかに記載のディスク制御装置。

【請求項７】

前記キャッシュメモリ部における障害を処理するためのパス制御機構が、前記キャッシュメモリ部の中にかえて、前記リソース管理部に備わっていることを特徴とする請求項３乃至６のいずれかに記載のディスク制御装置。

【請求項８】

前記障害監視機構により障害を検出した際、前記パス制御機構により前記障害部位の前記ローカルIDと前記障害部位の機能を引き継ぐ前記交換部位の前記ローカルIDとを交換し、該ローカルID交換に対応して前記スイッチ網内経路を切り換えるよう前記スイッチ内フォワーディングテーブルを変更することを特徴とする請求項１乃至７のいずれかに記載のディスク制御装置。

【請求項９】

ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、記憶装置とのインターフェースを有する複数のディスクインターフェース部と、前記記憶装置に対しリードまたはライトされるデータを一時的に格納する複数のキャッシュメモリ部を有し、各ホストインターフェース部は、前記ホストコンピュータとのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、各ディスクインターフェース部は、前記記憶装置とのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、

前記複数のホストインターフェース部と前記キャッシュメモリ部、及び、前記複数のディスクインターフェース部と前記キャッシュメモリ部との間が１つ以上のスイッチで構成されるスイッチ網を介して接続され、前記ホストインターフェース部、前記ディスクインターフェース部及び前記キャッシュメモリ部は、前記スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を有し、前記スイッチは前記スイッチ網内での経路を前記IDを用いて指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、さらに前記複数のキ

ャッシュメモリ部は、障害監視機構とパス制御機構を有するディスク制御装置における前記キャッシュメモリ部の障害処理方法であって、

前記障害監視機構が前記キャッシュメモリ部における障害発生の有無を確認するステップと、該障害監視機構が該障害発生を前記パス制御機構に通知するステップと、該パス制御機構が前記通知された障害情報を解析するステップと、該パス制御機構が障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するステップとを有することを特徴とする障害処理方法。

【請求項１０】

ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、記憶装置とのインターフェースを有する複数のディスクインターフェース部と、前記記憶装置に対しリードまたはライトされるデータを一時的に格納する複数のキャッシュメモリ部と、ホストインターフェース部及びディスクインターフェース部と前記キャッシュメモリ部との間のデータ転送に関する制御情報を格納する複数のリソース管理部を有し、前記複数のホストインターフェース部の各々は、前記ホストコンピュータとのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、前記複数のディスクインターフェース部の各々は、前記記憶装置とのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、

前記複数のホストインターフェース部、前記複数のディスクインターフェース部が１つ以上のスイッチで構成されるスイッチ網を介して前記キャッシュメモリ部と接続され、前記複数のホストインターフェース部、前記複数のディスクインターフェース部が前記スイッチ網を介して前記リソース管理部に接続され、前記ホストインターフェース部、前記ディスクインターフェース部、前記キャッシュメモリ部及び前記リソース管理部は、前記スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を有し、前記スイッチは前記スイッチ網内での経路を前記IDを用いて指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、さらに前記複数のキャッシュメモリ部と前記複数のリソース管理部は、障害監視機構とパス制御機構を有するディスク制御装置における前記キャッシュメモリ部および前記リソース管理部の障害処理方法であって、

前記障害監視機構が、前記キャッシュメモリ部または前記リソース管理部における障害発生の有無を確認するステップと、該障害監視機構が該障害発生を前記パス制御機構に通知するステップと、該パス制御機構が、前記通知された障害情報を解析するステップと、該パス制御機構が障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するステップを有することを特徴とする障害処理方法。

【請求項１１】

請求項１０記載のディスク制御装置の前記キャッシュメモリ部における障害を監視するための前記障害監視機構が、前記キャッシュメモリ部の中にかえて、前記リソース管理部の中にあるディスク制御装置における障害処理方法であって、該障害監視機構が前記キャッシュメモリ部の障害発生の有無を確認するステップが、前記リソース管理部の中で実行されることを特徴とする障害処理方法。

【請求項１２】

前記障害監視機構による障害発生の有無を確認するステップが、前記ホストインターフェース部及び前記ディスクインターフェース部でのリードまたはライト動作の際に実行されることを特徴とする請求項９乃至１１のいずれかに記載の障害処理方法。

【請求項１３】

請求項１０乃至１２のいずれかに記載のディスク制御装置の前記キャッシュメモリ部における障害を処理するためのパス制御機構が、前記キャッシュメモリ部の中にかえて、前記リソース管理部に備わっているディスク制御装置における障害処理方法であって、前記パス制御機構による前記フォワーディングテーブルを制御するステップが、前記リソース管理部の中で実行されることを特徴とする障害処理方法。

【請求項１４】

前記パス制御機構により前記フォワーディングテーブルを制御するステップは、前記障害部位のローカルIDと障害部位の機能を引き継ぐ交換部位のローカルIDとを交換するステップと、該ローカルID交換に対応して前記スイッチ網内経路を切り換えるよう前記スイッチ内フォワーディングテーブルを変更するステップを

有することを特徴とする請求項９乃至１３のいずれかに記載の障害処理方法。

【請求項１５】

前記記憶装置は磁気ディスク装置であることを特徴とする請求項１乃至８のいずれかに記載のディスク制御装置。

【請求項１６】

前記記憶装置は磁気ディスク装置であることを特徴とする請求項９乃至１４のいずれかに記載の障害処理方法。

【請求項１７】

複数のホストコンピュータと第１のネットワークを介して接続され、複数の磁気ディスク装置と第２のネットワークを介して接続されたディスク制御装置を備え、

上記ディスク制御装置は、

上記ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、

上記磁気ディスク装置とのインターフェースを有する複数のディスクインターフェース部と、

上記複数のホストインターフェース部及び上記複数のディスクインターフェース部との間が１つ以上のスイッチで構成されるスイッチ網を介して接続された複数のキャッシュメモリ部を備え、

前記複数のホストインターフェース部、前記複数のディスクインターフェース部及び前記複数のキャッシュメモリ部は、前記スイッチ網内で一意に決まるローカルＩＤと該ＩＤを変更する変更手段を有し、

前記スイッチは、前記スイッチ網内での経路を前記ＩＤを用いて指定するためのフォワーディングテーブルと、該テーブルを変更する変更手段を有し、

前記複数のキャッシュメモリ部は、

該複数のキャッシュメモリ部における障害発生の有無を監視するための障害監視機構と、

前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を有することを特徴とするディスクアレイサブシステム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、データを複数の磁気ディスク装置に格納するディスクシステム装置の制御装置に関する。

【０００２】

【従来の技術】

企業間の電子商取引や社会基盤としての金融システムなどでは、高度な信頼性が要求され、その根幹をなす基幹ストレージ・システムに対しても、極めて高い可用性が求められている。これら基幹ストレージ・システムでは、その可用性を高めるために、内部を冗長構成とし、障害発生時には自動的に故障個所を切り離して、正常な冗長部分で動作を継続するための自動障害回復機能を備えたディスク制御装置が広く使われている。

【０００３】

例えば、図９に示した、従来より知られているディスク制御装置は、ホストコンピュータ６０との間のデータ転送を実行する複数のホストインターフェース部１Xと、磁気ディスク装置７０との間のデータ転送を実行する複数のディスクインターフェース部２Xと、磁気ディスク装置７０のデータを一時的に格納するキャッシュメモリ部３Xと、ディスク制御装置１０４に関する制御情報（例えば、ホストインターフェース部１X及びディスクインターフェース部２Xとキャッシュメモリ部３Xとの間のデータ転送制御に関する情報、磁気ディスク装置７０に格納するデータの管理情報）を格納するリソース管理部５Xとを備えている。

ホストインターフェース部１X及びディスクインターフェース部２Xとキャッシュメモリ部３Xとの間は、データインターフェース信号６により接続される。ホストインターフェース部１Xとキャッシュメモリ部３Xとの間、及び、ディスクインターフェース部２Xとキャッシュメモリ部３Xとの間の接続にスイッチ４Xを用いることもある。ホストインターフェース部１X及びディスクインターフェース部２Xとリソース管理部５Xとの間は、管理インターフェース信号７により接続される。リソース管理部５Xと、ホストインターフェース部１X，及び、ディスク

インターフェース部２Xとの接続は、スイッチを介しても、介さなくても良い。これにより、リソース管理部５Xおよびキャッシュメモリ部３Xは全てのホストインターフェース部１X及びディスクインターフェース部２Xからアクセス可能な構成となる。

【０００４】

図１２に示すように、ホストインターフェース部１Xは、ホストインターフェース信号１との入出力を処理するチャネルプロトコル処理部９０、データインターフェース信号６との入出力を処理するための内部プロトコル処理部８X、管理インターフェース信号７との入出力を処理するためのプロセッサインターフェース１７、及びホストコンピュータ６０に対する入出力を制御するプロセッサ１４とローカルメモリ１５を有している。

ディスクインターフェース部２Xも、構造としては、ホストインターフェース部と同様であるが、ホストインターフェース信号１の代わりにディスクインターフェース信号２がチャネルプロトコル処理部９０に接続され、プロセッサ１４においては、ホストインターフェース部で行われる制御に加えて、RAID機能の実行も行う。

ホストインターフェース部１X、及び、ディスクインターフェース部２Xが、キャッシュメモリ部３Xと通信を行う場合、データの先頭に宛先アドレスを付加したパケットを使用してパケット転送を行う。

ホストインターフェース部１X、もしくは、ディスクインターフェース部２X内のプロセッサ１４の制御で生成されたパケットは、データインターフェース信号６を介して、スイッチ４Xに送られる。スイッチ４Xは、図１０で示したように、データインターフェース信号６に接続した複数のパスインターフェース４１Xと、パケットバッファ４３とアドレスラッチ４４とセレクタ４８を備えている。パスインターフェース４１X内には、パケットからアドレス情報を取り出すヘッダ解析部４２Xが含まれており、それによって解析抽出されたパケットアドレスがアドレスラッチ４４に取り込まれる。一方、送られてきたパケットは、パスインターフェース４１Xを通して、パケットバッファ４３に格納される。アドレスラッチ４４からは、パケット宛先に応じたセレクタ制御信号４７が生成され、パケ

ットバッファ４３に格納されたパケットの送出先をセレクタ４８によって切り換える。

【０００５】

スイッチ４Ｘで、宛先ごとに振り分けられたパケットは、再び、データインターフェース信号６を介して、目的のキャッシュメモリ部３Ｘに転送される。キャッシュメモリ部３Ｘは、図１１で示したように、データインターフェース信号６に接続した複数のデータパスインターフェース３１Ｘと、パケットバッファ３３と調停回路３９とセレクタ３８を備えている。データパスインターフェース３１Ｘ内には、パケットからアドレス情報を取り出すヘッダ解析部３２Ｘが含まれており、それによって解析抽出されたパケットアドレスは、調停回路３９に取り込まれる。一方、送られてきたパケットは、パスインターフェース３１Ｘを通して、パケットバッファ３３に格納される。調停回路３９は、複数のデータパスインターフェース３１Ｘの中から、いずれかを選択し、その選択結果に応じたセレクタ制御信号を生成する。このセレクタ制御信号で、セレクタ３８を切り換えることにより、所望のパケットバッファ３３の内容をキャッシュメモリ３７に、メモリ制御回路３５を介して、書き込むことができる。パケットバッファ３３に格納されたパケットがメモリ読み出しの要求であった場合、指定された領域のキャッシュメモリ３７の内容を、上記と逆の経路をたどることによって、ホストインターフェース部１Ｘ、もしくは、ディスクインターフェース部２Ｘに返送する。

【０００６】

ホストインターフェース部１Ｘ、及び、ディスクインターフェース部２Ｘが、リソース管理部５Ｘと通信を行う場合も、データインターフェース信号６の代わりに、管理インターフェース信号７が用いられる以外は、キャッシュメモリ部との通信と同様なパケット転送が行われる。リソース管理部５Ｘは、図１１に示したキャッシュメモリ部と、インターフェース信号を除いて同等な構成になっている。

ここで、キャッシュメモリ部３Ｘ、及び、リソース管理部５Ｘは、複数のホストインターフェース部１Ｘ、ディスクインターフェース部２Ｘからアクセスされるシステム共通のリソースであり、その可用性がシステムの信頼性に大きな影響を与え

るため、同等機能を複数備えた冗長構成となっており、たとえ、一方に障害が発生しても、残りの正常な部分で動作を継続できるように設計されている。具体的には、ホストインターフェース部１X、もしくは、ディスクインターフェース部２X内のいずれかのプロセッサ１４が、複数あるキャッシュメモリ部３X、もしくは、リソース管理部５Xのいずれかの障害を検知した場合、障害を検知したプロセッサ制御で、障害部分を閉塞して残りのキャッシュメモリ部３X、もしくは、リソース管理部５Xに、その機能を引き継がせるとともに、他の全てのプロセッサ１４に対して障害通知を行う。障害通知を受けた全てのプロセッサは、障害に伴うシステム構成・通信経路の変更処理をそれぞれ行うことによって、すべてのホストインターフェース部１Xとディスクインターフェース部２Xで、障害部分の切り離しを実現することが可能となる。

このように従来のディスク制御装置１０４では、キャッシュメモリ部３X、もしくは、リソース管理部５Xのような共通リソース障害に伴うシステム構成・通信経路の変更処理が、複数あるホストインターフェース１X、および、ディスクインターフェース２X内のプロセッサそれぞれで分散して行われていた。そのため、共通リソースの障害においては、分散配置されたプロセッサ間でのブロードキャスト通信を含む複雑な処理が必要であった。

【０００７】

ディスク制御装置を高信頼化する別の従来例として、共有システムリソースとシステムリソース・クライアントとの間で可用性の高いネットワーク通信を提供する障害処理機構が提案されている（例えば、特許文献１参照。）。この従来例においても、さきに説明した従来例と同様、複数あるプロセッサごとに経路変更（ルーティングテーブル変更）を行う方式となっている。

【０００８】

また、ディスク制御装置の高可用性を実現する別の従来例として、ホストコンピュータとディスクアレイサブセットとの間に配置して、両者の間のアドレス変換を行うスイッチを備えた記憶装置システムが提案されている（例えば、特許文献２参照。）。この従来例では、複数あるディスクアレイサブセットに障害が発生した場合、スイッチ内でパケットの解釈を行って障害部分への要求を同等の機

能を有した冗長部分の宛先に変更することで、経路変更などの障害処理を行う方式となっている。

【特許文献１】

特開２００２－４１３４８号公報

【特許文献２】

特開２０００－２４２４３４号公報

【０００９】

【発明が解決しようとする課題】

キャッシュメモリ部、リソース管理部のような共通リソースの障害は、ストレージ・システム、ひいては、ホストコンピュータで実行されているアプリケーションの動作不良を引き起こすため、すみやかな回復処理が行われなければならない。しかし、図９、図１０、図１１、図１２に示した従来技術では、すべてのホストインターフェース部１X、ディスクインターフェース部２Xにおいて経路変更が必要なため、障害処理に時間がかかり、ホストコンピュータとの間でのリード／ライトタスクの継続が行えずに、ストレージシステムの性能劣化やアプリケーション・プログラムの動作不良を引き起こす場合があった。また、この障害処理には、ホストインターフェース部１X、ディスクインターフェース部２Xに高機能プロセッサと複雑な制御プログラムが必要となり、製造コストの増大、信頼性の低下を招いていた。特許文献１に示した別の従来技術においても、複数あるプロセッサごとにルーティングテーブル変更を行う必要があり、同様の課題があった。

【００１０】

また、特許文献２で開示されている別の従来技術では、パケット宛先変更機能を有するスイッチを導入することによって、複数あるディスクアレイ・サブセット間で機能を引き継がせる等の障害回避動作をスイッチ内処理のみで行うことが可能となる。しかし、その一方で、パケットごとに宛先の解釈が必要となり、障害発生時のみならず、通常動作時においても、処理に多くの時間を要し、ストレージシステムの性能劣化を招くという課題を有している。

【００１１】

本発明の目的は、上記従来技術の欠点を改善し、障害発生時に、迅速かつ高信頼に障害処理を行い、かつ、通常動作時においても性能劣化を起こさない高可用のディスク制御装置、および、その障害処理方法を提供することにある。
より具体的には、本発明の目的は、システム共通リソースの障害時を含め、いかなる場合においても、ストレージ・システムの性能劣化や、ホスト・アプリケーションの動作不良を引き起こすことのない高可用性ディスク制御装置を提供することにある。

【００１２】

【課題を解決するための手段】

上記目的を達成するため、本発明では、ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、磁気ディスク装置とのインターフェースを有する複数のディスクインターフェース部と、磁気ディスク装置に対しリード／ライトされるデータを一時的に格納するキャッシュメモリ部と、ホストインターフェース部及びディスクインターフェース部とキャッシュメモリ部との間のデータ転送に関する制御情報を格納するリソース管理部を有し、各ホストインターフェース部は、ホストコンピュータとのインターフェースとキャッシュメモリ部との間のデータ転送を実行し、各ディスクインターフェース部は、磁気ディスク装置とのインターフェースとキャッシュメモリ部との間のデータ転送を実行するディスク制御装置であって、

複数のホストインターフェース部とキャッシュメモリ部及び複数のディスクインターフェース部とキャッシュメモリ部との間が１つ以上のスイッチで構成されるスイッチ網を介して接続され、スイッチがスイッチ網内での経路を指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、また、ホストインターフェース部、ディスクインターフェース部及びキャッシュメモリ部が、スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を有し、さらに複数のキャッシュメモリ部が、その障害発生の有無を相互に監視するための障害監視機構と障害発生時に障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を有するディスク制御装置を提供する。

【００１３】

【発明の実施の形態】

以下、大容量のデータの記憶装置として磁気ディスク装置を例にとって説明するが、大容量記憶装置として磁気ディスクに限られるものではなく、例えばＤＶＤのような大容量記憶装置であって良い。

本発明の実施の形態の１つとして、好ましくは、前記複数のホストインターフェース部と前記キャッシュメモリ部との間、及び前記複数のディスクインターフェース部と前記キャッシュメモリ部との間を、１つ以上のスイッチで構成されるスイッチ網を介して接続し、スイッチにスイッチ網内での経路を指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を設け、また、ホストインターフェース部、ディスクインターフェース部及びキャッシュメモリ部に、スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を設け、さらに複数のキャッシュメモリ部に、その障害発生の有無を監視するための障害監視機構と障害発生時に障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を設ける。

また、好ましくは、前記複数のホストインターフェース部、前記複数のディスクインターフェース部、前記キャッシュメモリ部、前記リソース管理部との間を１つ以上のスイッチで構成されるスイッチ網を介して接続し、スイッチにスイッチ網内での経路を指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を設け、また、ホストインターフェース部、ディスクインターフェース部及びキャッシュメモリ部に、スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を設け、さらに複数のキャッシュメモリ部に、その障害発生の有無を監視するための障害監視機構と障害発生時に障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を設ける。

また、好ましくは、前記リソース管理部に、前記キャッシュメモリ部、もしくは、前記リソース管理部の障害を監視するための障害監視機構を設ける。

【００１４】

また、好ましくは、前記ホストインターフェース部及び前記ディスクインター

フェース部に前記キャッシュメモリ部、もしくは、前記リソース管理部の障害監視機構に対して障害を報告する機能を設ける。

また、好ましくは、前記リソース管理部に、前記キャッシュメモリ部の障害を処理するためのパス制御機構を設ける。

その他、本願が開示する課題、及びその解決方法は、発明の実施形態の欄及び図面により明らかにされる。

【００１５】

以下、本発明の実施例を図面を用いて説明する。

《実施例１》

図１、図２、図６、図７、及び図８に、本発明の一実施例を示す。

図２に示したディスク制御装置１００は、ホストコンピュータ６０とのインターフェース部（ホストインターフェース部）１０と、磁気ディスク装置７０とのインターフェース部（ディスクインターフェース部）２０と、キャッシュメモリ部３０、スイッチ４０、リソース管理部５０を有し、ホストインターフェース部１０及びディスクインターフェース部２０と、キャッシュメモリ部３０及びリソース管理部５０との間は、スイッチ４０を介して、内部インターフェース信号４で接続されている。すなわち、内部インターフェース信号４を介して、全てのホストインターフェース部１０、全てのディスクインターフェース部２０から、全てのキャッシュメモリ部３０、あるいはリソース管理部５０へアクセス可能な構成となっている。

【００１６】

図８に示すように、ホストインターフェース部１０は、ホストインターフェース信号１との入出力を処理するチャネルプロトコル処理部９０、データインターフェース信号との入出力を処理するための内部プロトコル処理部８０を有し、ホストコンピュータ６０とキャッシュメモリ部３０間のデータの転送、及びリソース管理部５０との間の制御情報の転送を実行する。

ディスクインターフェース部２０も、構造としては、ホストインターフェース部と同様であるが、ホストインターフェース信号１の代わりにディスクインターフェース信号２を有し、磁気ディスク装置７０とキャッシュメモリ部３０間のデ

ータの転送、及びリソース管理部５０との間の制御情報の転送を実行する。

キャッシュメモリ部３０は、図７に示すように、内部インターフェース信号４との入出力処理を行う内部プロトコル処理部８０とプロセッサ３６とキャッシュメモリ３７とメモリ制御回路３５とＤＭＡエンジン３４を有し、磁気ディスク装置７０へ記録するデータや磁気ディスク装置から読み出したデータを一時的に格納する。

【００１７】

リソース管理部５０も、キャッシュメモリ部３０と同等の構成を有し、システム構成などの管理制御情報を維持する。

【００１８】

スイッチ４０は、図６に示すように、内部インターフェース信号４に接続した複数のパスインターフェース４１と、パケットバッファ４３とアドレスラッチ４４とセレクタ４８を備え、ホストインターフェース部１０およびディスクインターフェース部２０と、キャッシュメモリ部３０およびリソース管理部４０との間の経路接続を行う。

なお、可用性向上のために、ホストインタフェース部１０、ディスクインターフェース部２０、キャッシュメモリ部３０、リソース管理部５０を、それぞれ複数のポートを有する構成とし、スイッチ４０との間にそれぞれ複数の転送経路を設けることも出来る。

【００１９】

ホストインターフェース部１０、ディスクインターフェース部２０、キャッシュメモリ部３０、リソース管理部５０の内部プロトコル処理部８０には、内部インターフェース信号４の接続しているスイッチ網内での宛先を一意に特定するためのローカルＩＤ（ＬＩＤ）を保持するＬＩＤ情報８１が備わっている。

【００２０】

一方、スイッチ４０内には、ポート番号（パスインターフェース４１の位置）とＬＩＤとの対応関係を示すフォワーディング・テーブル４６が備わっている。フォワーディング・テーブル４６の例を、図１（１）に示す。この例では、２つのホストインターフェース１０と２つのディスクインターフェース２０が２つの

スイッチ４０A、４０Bを介して、２つのキャッシュメモリ（共通リソース）３０A、３０Bに接続している。ホストインターフェース１０、ディスクインターフェース２０、キャッシュメモリ３０A、３０Bはそれぞれ２つずつの内部インターフェース信号とそれに対応したローカルＩＤ（ＬＩＤ）情報を持っている。スイッチ４０A、４０Bは、それぞれ８つのポート（パスインターフェース４１）とそれに対応したポート番号を持っている。フォワーディング・テーブル４６とは、このＬＩＤとポート番号との対応表であって、例えば、スイッチ４６Aのフォワーディング・テーブルＡでは、ＬＩＤ①、③、⑤、⑦、⑨、(11)が、それぞれポートａ、ｂ、ｃ、ｄ、ｅ、ｆと接続していることを示している。このフォワーディング・テーブルを参照することにより、パケット宛先（ＬＩＤ）によって、どのポートに送出すればよいかがわかる。

【００２１】

内部インターフェース信号の接続しているスイッチ網は、例えば、キャッシュメモリ部３０内のプロセッサ３６で実行されるネットワーク管理プログラムによって、維持管理されている。ネットワーク内のＬＩＤ情報８１やスイッチ内のフォワーディング・テーブル４６は、内部インターフェース信号４を介して、ネットワーク管理プログラムにより、設定更新される。

本発明のディスク制御装置における通常動作の一例として、図２、図６、図７、図８を用い、ホストコンピュータ６０からディスク制御装置１００を介して磁気ディスク装置７０へ読み出し要求を発行する場合の動作について説明する。

まず、ホストコンピュータ６０は、自身が接続されているホストインターフェース部１０にデータの読出し要求を発行する。 要求を受けたホストインターフェース部１０は、リソース管理部５０にアクセスし、要求されたデータがどの磁気ディスク装置７０内に格納されており、当該磁気ディスクがどのキャッシュメモリ部３０に制御されているかを調べる。リソース管理部５０には、要求データのアドレスからこれらの情報を検索するためのテーブルが格納されており、要求されたデータをもとに管轄のキャッシュメモリ部を調べることができる。次に、要求を受けたホストインターフェース部１０では、要求データを管理しているキャッシュメモリ部３０に、読み出し要求を転送する。キャッシュメモリ部３０は

、キャッシュメモリ３７に要求されたデータが格納されているかどうかを確認する。キャッシュメモリ部３０内にデータが存在しなかった場合、プロセッサ３６は、要求データを磁気ディスク装置７０から読出し、キャッシュメモリ３７に格納する。キャッシュメモリ部３０は、キャッシュメモリ３７内に格納された要求データを、ホストインターフェース部１０まで転送し、ホストコンピュータ６０に送る。

ホストインターフェース部１０及びディスクインターフェース部２０がキャッシュメモリ部３０、及びリソース管理部５０とスイッチ４０を介して通信する場合、パケット宛先にＬＩＤ情報を用い、スイッチでの経路変更にフォワーディング・テーブル４６を使用する。

例えば、ホストコンピュータ６０からの読み出し要求を、スイッチ４０を介して、キャッシュメモリ部３０に転送する場合、ホストインターフェース部１０では、ホストインターフェース信号１を、チャネルプロトコル処理部９０と内部プロトコル処理部８０を介して、内部インターフェース信号４に変換する。その際、読み出し要求パケットの宛先には、送り先のキャッシュメモリ部のＬＩＤを設定する。ホストインターフェース部からの要求パケットを受けたスイッチ４０では、経路変更手段４５内のフォワーディング・テーブル４６に従い、ヘッダ解析部４２で抽出されたパケット宛先に応じたセレクタ制御信号４７を生成し、パケットバッファ４３に格納されたパケットの送出先をセレクタ４８によって切り換えて、所望のキャッシュメモリ部３０に向けてパケットを転送する（図16）。

【００２２】

ディスクインターフェース部２０が、キャッシュメモリ部と通信を行う場合、もしくは、ホストインターフェース部１０、及び、ディスクインターフェース部２０が、リソース管理部５０と通信を行う場合も、ホストインターフェース部とキャッシュメモリ部との間の通信と同様なパケット転送が行われる。

なお、リソース管理部５０と、ホストインターフェース部１０およびディスクインターフェース部２０との接続に使われているスイッチ４０を含んだスイッチ網は、キャッシュメモリ部３０とホストインターフェース部１０およびディスクインターフェース部２０との接続に使われているスイッチ網と同じであっても、別

の専用ネットワークであっても良い（図17）。また、スイッチ４０を使わずに、直接接続する形態であっても構わない。

次に、本発明のディスク制御装置の特徴である障害回復動作の一例として、図１、図２を用い、２つのキャッシュメモリ部３０間での障害監視機構とパス制御機構の動作について説明する。

キャッシュメモリ部３０は、その可用性向上のため、同等機能を有するマスタとスレーブの２つのキャッシュメモリ部を有している。スレーブキャッシュメモリは、マスタキャッシュメモリに障害が発生した際に、その機能を引き継ぐためのもので、ホットスタンバイで動作している。これらマスタ・キャッシュメモリ部とスレーブ・キャッシュメモリ部は、スイッチ４０を介して、相互にその動作を確認するための障害監視機構Ｃを有している。すなわち、一定間隔ごとに自分の動作状況を報告するパケットを生成し、お互い相手の動作状態を監視することができる。その動作の概要を図１３に示す。障害監視機構では通信が行われる度に、正常なシーケンスとＡＣＫがチェックされ、異常が発見された場合には、主経路もしくは副経路を介して、パス制御機構に障害通知を行う。図２に示した構成では、マスタおよびスレーブキャッシュメモリ部が、それぞれパス制御機構を有しているので、そのいずれかのキャッシュメモリ部に異常が発生した場合、その障害情報は、ただちにもう一方のキャッシュメモリ部で検知することができる。障害を検知したキャッシュメモリ部では、障害を起こしているキャッシュメモリ部を閉塞するとともに、内部に備えたパス制御機構Ｐによって、システム構成を変更し、ホストインターフェース部１０およびディスクインターフェース部２０が、障害キャッシュメモリ部にアクセスしないように制御する。

以下、パス制御機構Ｐの仕組みを、図１４を使って説明する。障害監視機構から障害通知を受けたパス制御機構は、その通知の妥当性を確認した後、可用性向上のため複数個設けてあるパス制御機構間で、障害情報の同期化を行う。その後、障害解析を行い、現時点で障害特定が可能かどうか判断する。特定が出来ない場合、特定可能になるまで障害処理を遅延させる。特定可能になった場合、障害通知情報によりアクセス障害なのか機能障害なのかを判断し、機能障害の場合、障

害部位を含んだ冗長部位間で、処理途中のＪＯＢなどの同期化処理を試みる。その後、障害部位に対するアクセス・パスを冗長部位に振り替える交替パス処理を行う。図１で具体的に説明すると、マスタキャッシュメモリ部で機能障害が発生しスレーブキャッシュメモリ部にフェールオーバする場合は、以下のようになる。障害のない正常動作状態における、ホストインターフェース部１０、ディスクインターフェース部２０、キャッシュメモリ部３０Ａ、３０Ｂ、スイッチ４０Ａ、４０ＢのＬＩＤとフォワーディング・テーブルの値は、図１（１）に示した通りである。ここで、マスタ・キャッシュメモリ部３０Ａに障害が発生し、スレーブ・キャッシュメモリ部３０Ｂの障害監視機構によって、当該障害が検出された場合、３０Ｂのパス制御機構Ｐにより、３０Ａの機能を３０Ｂが引き継ぐとともに、３０Ａ宛のパケットを３０Ｂに振り替える制御を行う。すなわち、３０Ａの２つのＬＩＤ(⑨、(10))と３０Ｂの２つのＬＩＤ((11)、(12))とを交換し、それに対応して、フォワーディング・テーブル４６Ａと４６Ｂを変更する。その結果、ＬＩＤとフォワーディングテーブルは図１（２）のようになり、３０Ａへのアクセスは、すべて３０Ｂに振り向けられて、障害部位３０Ａのシステムからの切り離しが完了する。３０Ａの動作を３０Ｂが引き継ぐには、３０Ａと３０Ｂの内容が一致している必要があるが、それらは、通常の同期動作により実現される。すなわち、常に３０Ａと３０Ｂが同じ内容になるよう、両方に対して同じアクセス要求を発生させる方法や、定期的に両者でデータをコピーする方法が考えられる。

なお、リソース管理部５０も同様の障害監視機構Ｃとパス制御機構Ｐを備え、同様の手順で障害回復動作を行うことができる。これら障害監視機構やパス制御機構は、キャッシュメモリ部３０、もしくは、リソース管理部５０内のプロセッサ３６によって実行される制御プログラムとして実現することができる。また、本実施例では、リソース管理部５０を独立に設けた構成としているが、リソース管理部をキャッシュメモリ部３０の中に設けることもできる。また、図１で、スレーブキャッシュメモリ部内のＬＩＤとマスタキャッシュメモリ部内のＬＩＤを交換するのではなく、スレーブ側でマスタ側のＬＩＤを追加して持つようにすることも可能である。その場合、障害前のスレーブ側ＬＩＤが障害後も有効になると

いうメリットがある。

【００２３】

本実施例によれば、キャッシュメモリ部３０もしくはリソース管理部５０の障害に対して、スイッチ４０のフォワーディング・テーブルとキャッシュメモリ部３０もしくはリソース管理部５０のＬＩＤの変更のみで、障害部位の切り離しが完了し、従来技術のように、複数のホストインターフェース部１０及びディスクインターフェース部２０の間でのブロードキャスト通信や複雑な制御は不要である。そのため、障害発生時においても、迅速、かつ、高信頼な障害回復処理を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがない。

また、本実施例のスイッチ内フォワーディング・テーブルは、障害発生時にのみ変更されものであって、従来技術のように、通信を行うたびにパケットの宛先の解釈や変更が行われる複雑なスイッチを必要としない。そのため、障害のない通常の動作においての性能劣化は皆無で、かつ、低コストで高信頼に製造することが可能である。

【００２４】

《実施例２》

図３に、本発明の他の実施例を示す。

図３に示した実施例は、キャッシュメモリ部３０、及び、リソース管理部５０が、それぞれ障害通知のための専用線であるハートビート信号３を備えていることと、内部インターフェース信号のスイッチ網が多段のスイッチ４０で構成されていることを除いて、実施例１の図２に示す構成と同様である。ただし、キャッシュメモリ部３０は、障害監視機構Ｃのみを備え、パス制御機構は実装されていない。リソース管理部５０は、障害監視機構Ｃ、パス制御機構Ｐともに備えている。

キャッシュメモリ部、および、リソース管理部は、マスタとスレーブにより冗長構成を形成しており、基本的には同じデータ内容を保持している。ただし、キャッシュメモリ部におけるディスクからの読み出しデータについては、マスタとスレーブ間で同一の内容を保持していなくても構わない。

障害が発生した場合の処理は、基本的に、実施例１と同様であるが、その概要を図１８で簡単に説明する。マスタ／スレーブ キャッシュメモリ部、および、マスタ／スレーブリソース管理部は、障害監視機構により、相互に障害発生の有無について定期的な確認を行っている。障害監視機構によって発見された障害は、リソース管理部のパス制御機構に通知される。パス制御機構では、通知された障害情報の解析を行い、障害部位の特定を行う。パス制御機構では、障害部位が特定できれば、スイッチ内にあるフォワーディングテーブルを制御することで、障害部位を回避するようにパスの設定を行うことで、障害部位の切り離しが完了する。

本実施例では、専用のハートビート信号３を設けることにより、相互にその動作を確認するための障害監視機構Ｃを、実施例１より単純な構成で実現することができる。すなわち、ハートビート信号３を通して、お互いの動作状況を、直接、監視することができる。そのため、マスタ、もしくは、スレーブのいずれかのキャッシュメモリ部、もしくはリソース管理部に異常が発生した場合、その障害情報は、より迅速にもう一方のキャッシュメモリ部、もしくは、リソース管理部で検知することができる。

また、本実施例では、キャッシュメモリ部３０内で検出された障害情報が、スイッチ４０を介して(マスタ)リソース管理部５０のパス制御機構Ｐに通知され、リソース管理部内のパス制御機構Ｐによってキャッシュメモリ部３０の障害回復処理が行われる。これにより、障害情報をリソース管理部５０に集めて、より適切な障害回復処理を行うことが可能となる。

また、本実施例では、ホストインターフェース側とディスクインターフェース側でスイッチを分離することにより、ホスト側とディスク側で柔軟な接続数変更が可能になるとともに、より大規模な構成への対応が可能となる。

【００２５】

本実施例によれば、前述の実施例と同様、迅速、かつ、高信頼な障害回復処理を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがなく、また、障害のない通常の動作においての性能劣化は皆無で、かつ、低コストで高信頼に製造することが

可能である。

【００２６】

《実施例３》

図４に、本発明の他の実施例を示す。

図４に示した実施例は、キャッシュメモリ部３０、及び、リソース管理部５０が、ハートビート信号３を備えていないことと、内部インターフェース信号のスイッチ網が冗長構成になっていることを除いて、実施例２の図３に示す構成と同様である。ただし、キャッシュメモリ部３０は、障害監視機構Ｃとパス制御機構を実装していない。リソース管理部５０は、障害監視機構Ｃ、パス制御機構Ｐともに備えている。

本実施例では、キャッシュメモリ部３０の障害監視も、リソース管理部５０内の障害監視機構Ｃを用いて行う。その実現方法としては、リソース管理部の障害監視機構Ｃが、定期的にキャッシュメモリ部３０へアクセスを行い、キャッシュメモリ部の動作状況を監視する方法や、ホストインターフェース部１０およびディスクインターフェース部２０からキャッシュメモリ部３０へアクセスした際に障害が検出された場合、その障害情報をリソース管理部に報告する方法などが考えられる。また、本実施例では、各ホストインターフェース部、各ディスクインターフェース部に複数のポートを設け、スイッチも２重化することにより、ホストインターフェース部、および、ディスクインターフェース部とキャッシュメモリ部、もしくは、リソース管理部との間に複数のパスを設けてある。

これにより、リソース管理部、キャッシュメモリ部の機能障害だけでなく、リソース管理部、もしくは、キャッシュメモリ部とホストインターフェース部およびディスクインターフェース部との間のパス障害についても、障害回復を行えることになり、可用性をさらに高めることができる。

また、リソース管理部５０に、障害監視機構とパス制御機構を集めることにより、より障害状況の的確な分析が行えることになり、適切かつ高信頼の障害回復処理が可能となる。

【００２７】

本実施例によれば、前述の実施例と同様、迅速、かつ、高信頼な障害回復処理

を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがなく、また、障害のない通常の動作においての性能劣化は皆無で、かつ、低コストで高信頼に製造することが可能である。

【００２８】

《実施例４》

図５に、本発明の他の実施例を示す。

図５に示した実施例は、キャッシュメモリ部３０、及び、リソース管理部５０が、ハートビート信号３を備えていないことと、複数のディスク制御サブユニット２００を備えることを除いて、実施例２の図３に示す構成と同様である。複数のディスク制御サブユニット内のキャッシュメモリ部それぞれが、障害監視機構Ｃを備えている。リソース管理部５０は、障害監視機構Ｃ、パス制御機構Ｐともに備えている。

本実施例では、ディスク制御サブユニット２００ごとに、キャッシュを分散配置することで、キャッシュの使用効率（ヒット率）を高めて性能を向上するとともに、ホスト側とディスク側で柔軟にシステム規模を拡張することが可能となり、より高スケーラブルなシステムの提供が可能となる。

また、本実施例では、実施例２と同様、キャッシュメモリ部３０の障害に伴う障害回復処理も、リソース管理部５０内のパス制御機構Ｐを用いて行う。実施例２、３と同様、リソース管理部５０に、障害情報を集めることにより、より障害状況の的確な分析が行えることになり、ディスク制御サブユニット２００の個数を増加させた、より大規模なディスク制御装置においても、適切かつ高信頼な障害回復処理が可能となる。

【００２９】

本実施例によれば、前述の実施例と同様、迅速、かつ、高信頼な障害回復処理を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがなく、また、障害のない通常の動作においての性能劣化は皆無で、かつ、低コストで高信頼に製造することが可能である。

【００３０】

《実施例５》

図１５に、本発明の他の実施例を示す。

図１５に示した実施例では、実施例１－４で記載のディスク制御装置が、複数のホストコンピュータとホストコンピュータ間ネットワークを介して接続し、複数の磁気ディスク装置と磁気ディスク装置間ネットワークを介して接続している。ホストコンピュータ間ネットワークには、ファイルシステムの処理を行うサーバ１１０（NASヘッド）、複数のディスク制御装置が管轄するストレージをまとめて管理するためのサーバ１２０（ディスク制御装置間仮想化エンジン）、データベース・インターフェース処理を行うためのサーバ１３０（DB機能付加エンジン）などが接続されることもある。これら、NASヘッド、仮想化エンジン、DB機能付加エンジンは、ディスク制御装置の中に実装しても構わない。

【００３１】

本実施例によれば、迅速、かつ、高信頼な障害回復処理を行えるディスク制御装置を用いることにより、極めて可用性が高く、性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことのない、ストレージシステムの提供が可能となる。

【００３２】

【発明の効果】

以上説明したように、本発明では、キャッシュメモリ部３０もしくはリソース管理部５０の障害に対して、スイッチ４０のフォワーディング・テーブルとキャッシュメモリ部３０もしくはリソース管理部５０のＬＩＤの変更のみで、障害部位の切り離しが完了し、従来技術のように、複数のホストインターフェース部１０及びディスクインターフェース部２０の間でのブロードキャスト通信や複雑な制御は不要である。そのため、障害発生時においても、迅速、かつ、高信頼な障害回復処理を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがない。

また、本発明において、スイッチ内フォワーディング・テーブルは、障害発生時にのみ変更されものであって、従来技術のように、通信を行うたびにパケット

の宛先の解釈や変更が行われる複雑なスイッチを必要としない。そのため、障害のない通常の動作においての性能劣化は皆無で、かつ、低コストで高信頼に製造することが可能である。

【００３３】

本発明では、障害監視機構による障害発生通知を、パス制御機構で解析し、フォワーディングテーブルを制御するので、柔軟なシステム構成に対応可能である。とくに、複数のディスク制御サブユニットが存在する大規模なディスク制御装置の場合でも、複数の障害監視機構からの障害情報をパス制御機構に集めることにより、より障害状況の的確な分析が行え、高信頼な障害回復処理が可能となる

【図面の簡単な説明】

【図１】

本発明によるディスク制御装置における障害回復処理の動作原理を示す図である。

【図２】

本発明によるディスク制御装置の構成を示す図である。

【図３】

本発明によるディスク制御装置の構成を示す図である。

【図４】

本発明によるディスク制御装置の構成を示す図である。

【図５】

本発明によるディスク制御装置の構成を示す図である。

【図６】

本発明によるディスク制御装置内のスイッチの構成を示す図である。

【図７】

本発明によるディスク制御装置内のキャッシュメモリ部の構成を示す図である。

【図８】

本発明によるディスク制御装置内のホストインターフェース部の構成を示す図である。

【図９】

従来のディスク制御装置の構成を示す図である。

【図１０】

従来のディスク制御装置内のスイッチの構成を示す図である。

【図１１】

従来のディスク制御装置内のキャッシュメモリ部の構成を示す図である。

【図１２】

従来のディスク制御装置内のホストインターフェース部の構成を示す図である。

【図１３】

本発明の障害監視機構の動作を示す図である。

【図１４】

本発明のパス制御機構の動作を示す図である。

【図１５】

本発明のディスク制御装置を用いたストレージシステムの例を示す図である。

【図１６】

本発明のディスク制御装置においてホストコンピュータからキャッシュメモリ部へのコマンド送信の概要を示す図である。

【図１７】

本発明のディスク制御装置の構成を示す図である。

【図１８】

本発明のディスク制御装置の障害処理の概要を示す図である。

【符号の説明】

１　ホストインターフェース信号

２　ディスクインターフェース信号

３、５　ハートビート信号

４　内部インターフェース信号

６　データインターフェース信号

７　管理インターフェース信号

１０、１Ｘ　ホストインターフェース部

１１　ＤＭＡエンジン

１２　送信データインターフェース

１３　受信データインターフェース

１４　プロトコル制御プロセッサ

１５　ローカルメモリ

１６　バスインターフェース

１７　プロセッサインターフェース

２０、２Ｘ　ディスクインターフェース部

３０、３０Ａ、３０Ｂ、３Ｘ　キャッシュメモリ部

３１Ｘ　データパスインターフェース

３２Ｘ　ヘッダ解析部

３３　パケットバッファ

３４　ＤＭＡエンジン

３５　メモリ制御回路

３６　プロセッサ

３７　キャッシュメモリ

３８　セレクタ

３９　調停回路

４０、４０Ａ、４０Ｂ、４Ｘ　スイッチ

４１　パスインターフェース（ＰＯＲＴ）

４１Ｘ　データパスインターフェース

４２、４２Ｘ　ヘッダ解析部

４３　パケットバッファ

４４　アドレスラッチ

４５　経路変更手段

４６、４６Ａ、４６Ｂ　フォワーディング・テーブル

４７　セレクタ制御信号

４８　セレクタ

５０　リソース管理部

６０　ホストコンピュータ

７０　磁気ディスク装置

８０、８Ｘ　内部プロトコル処理部

８１　ローカルＩＤ情報（ＬＩＤ）

８２　送信ＰＨＹ

８３　受信ＰＨＹ

８４　プロトコル制御プロセッサ

８５　ローカルメモリ

８６　バッファ

８７　リンク処理

８８　ヘッダ制御

９０　チャネルプロトコル処理部

９２　送信ＰＨＹ

９３　受信ＰＨＹ

９６　バッファ

９７　リンク処理

９８　トランスポート処理

１００～１０４　ディスク制御装置

１１０　ＮＡＳヘッド

１２０　ディスク制御装置間仮想化エンジン

１３０　ＤＢ機能付加エンジン

２００　ディスク制御サブユニット

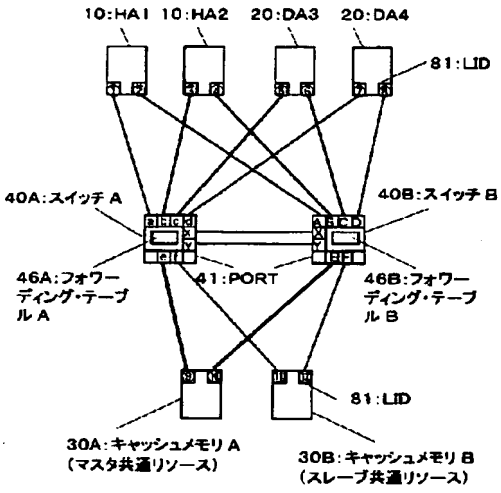
Ｃ　障害監視機構

Ｐ　パス制御機構。

【書類名】　図面

【図１】

図1



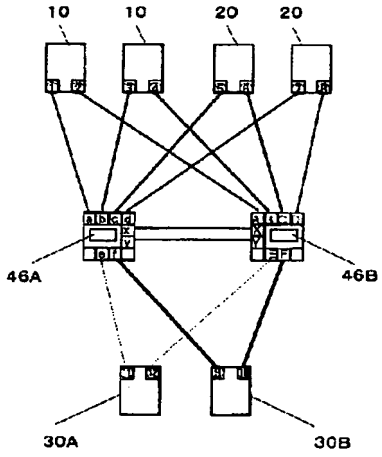
フォワーディング・テーブル A

LID	PORT
①	a
②	x
③	b
④	x
⑤	c
⑥	y
⑦	d
⑧	y
⑨	e
⑩	x
⑪	f
⑫	y

フォワーディング・テーブル B

LID	PORT
①	X
②	A
③	X
④	B
⑤	Y
⑥	C
⑦	Y
⑧	D
⑨	X
⑩	E
⑪	Y
⑫	F

（1）障害前の状態



フォワーディング・テーブル A

LID	PORT
①	a
②	x
③	b
④	x
⑤	c
⑥	y
⑦	d
⑧	y
⑨	f
⑩	x
⑪	–
⑫	–

フォワーディング・テーブル B

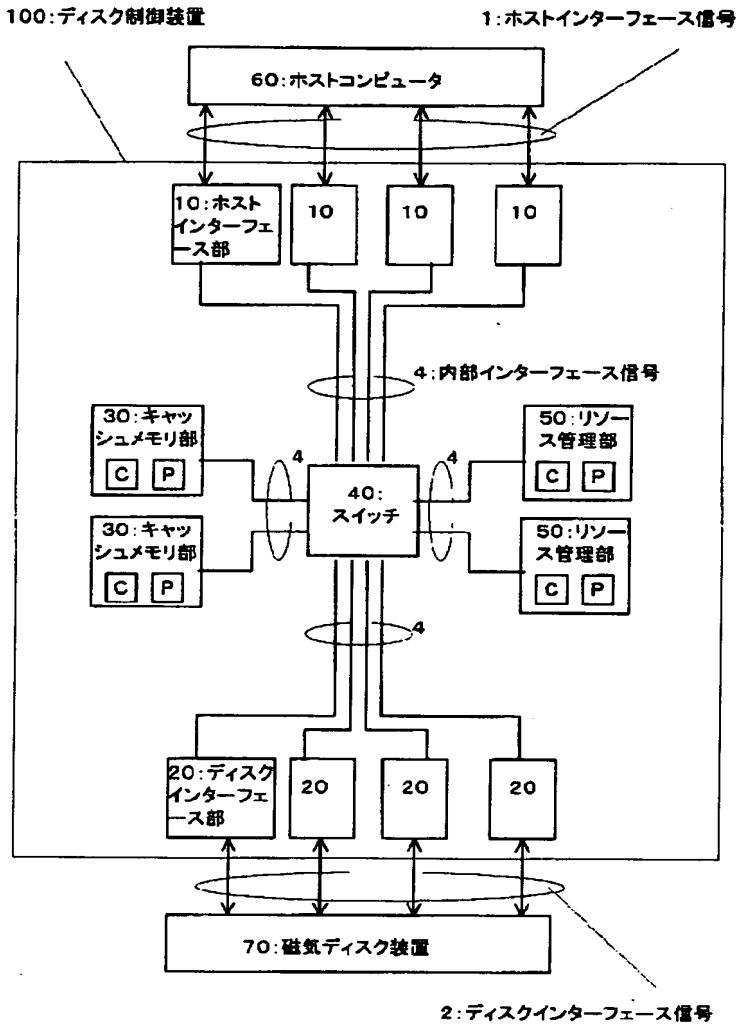
LID	PORT
①	X
②	A
③	X
④	B
⑤	Y
⑥	C
⑦	Y
⑧	D
⑨	X
⑩	F
⑪	–
⑫	–

（2）30Aの障害処理後の状態

出証特２００３－３０３５５９１

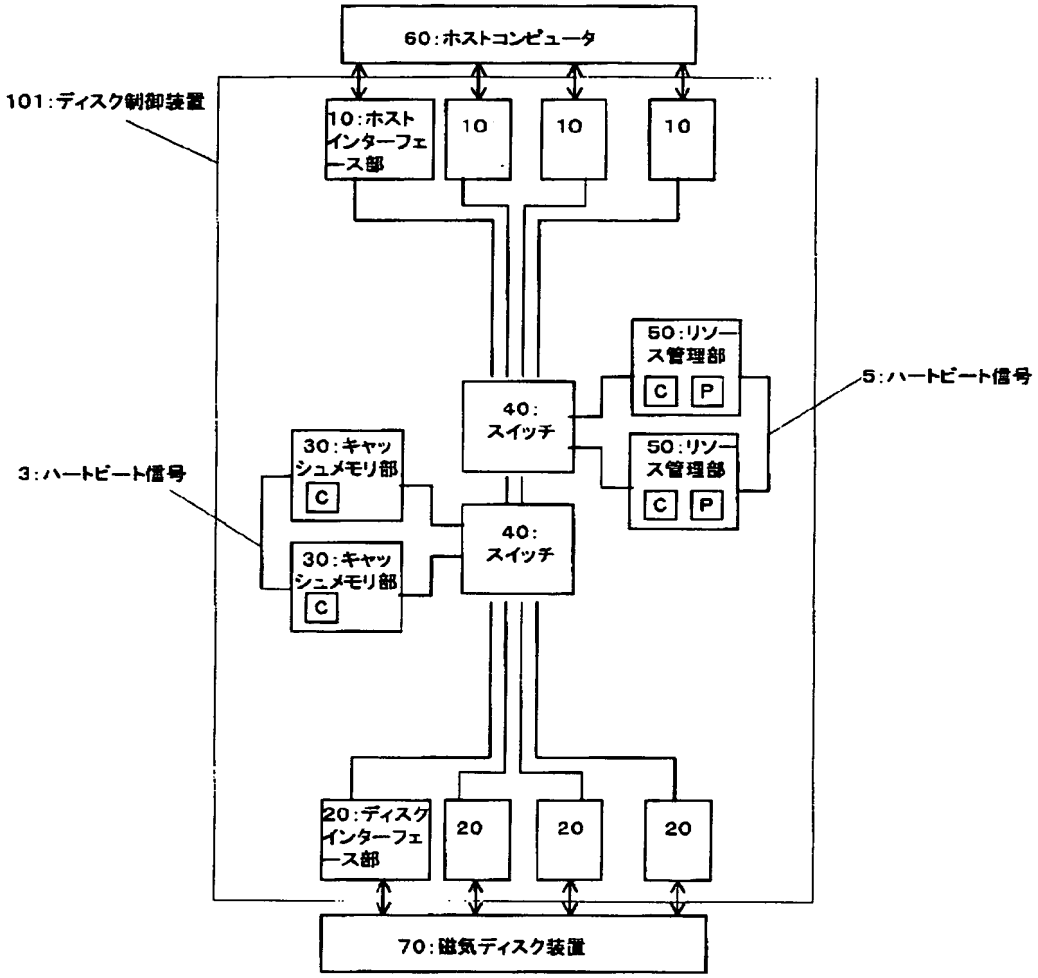
【図２】

図2


100:ディスク制御装置
1:ホストインターフェース信号
60:ホストコンピュータ
10:ホストインターフェース部
10
10
10
4:内部インターフェース信号
30:キャッシュメモリ部
C
P
30:キャッシュメモリ部
C
P
4
4
40:スイッチ
50:リソース管理部
C
P
50:リソース管理部
C
P
4
20:ディスクインターフェース部
20
20
20
70:磁気ディスク装置
2:ディスクインターフェース信号

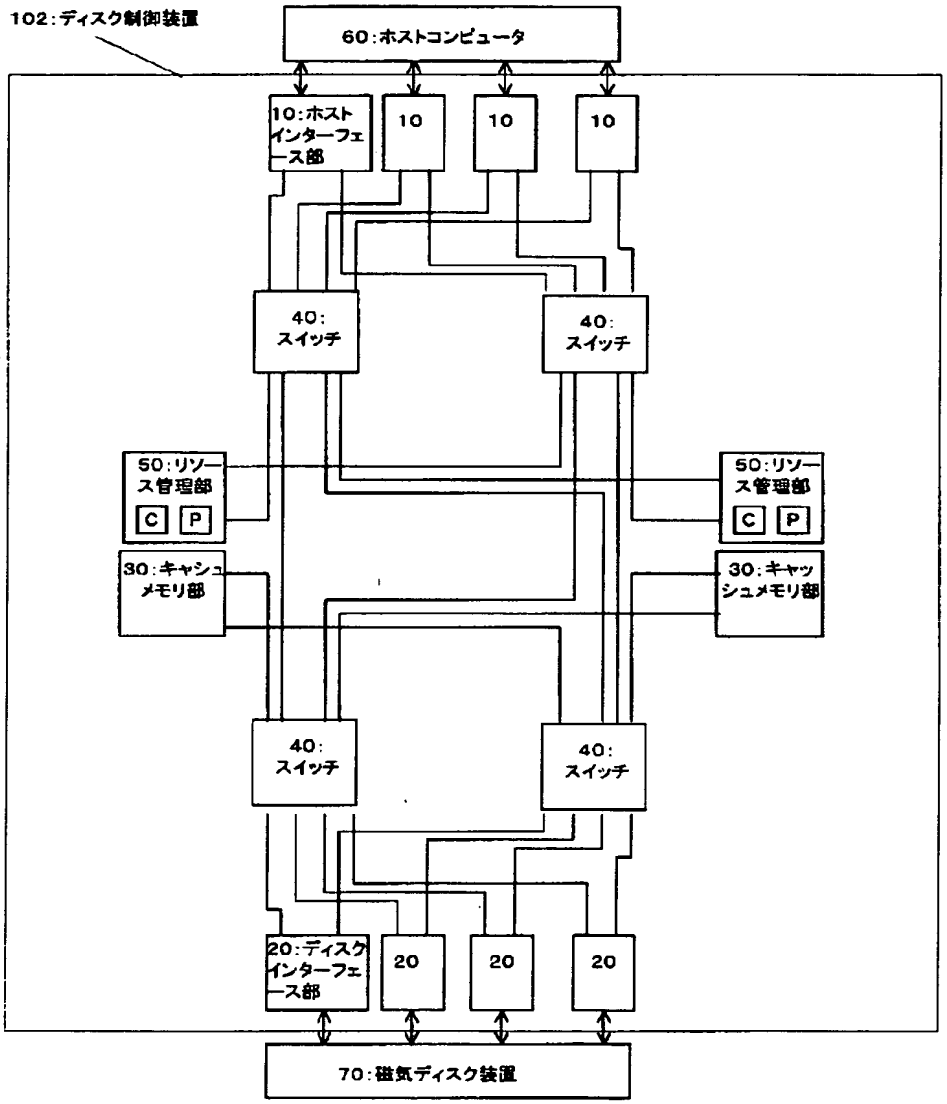
【図３】

図3


60:ホストコンピュータ
101:ディスク制御装置
10:ホストインターフェース部
10
10
10
50:リソース管理部
C
P
40:スイッチ
5:ハートビート信号
3:ハートビート信号
30:キャッシュメモリ部
C
50:リソース管理部
C
P
40:スイッチ
30:キャッシュメモリ部
C
20:ディスクインターフェース部
20
20
20
70:磁気ディスク装置

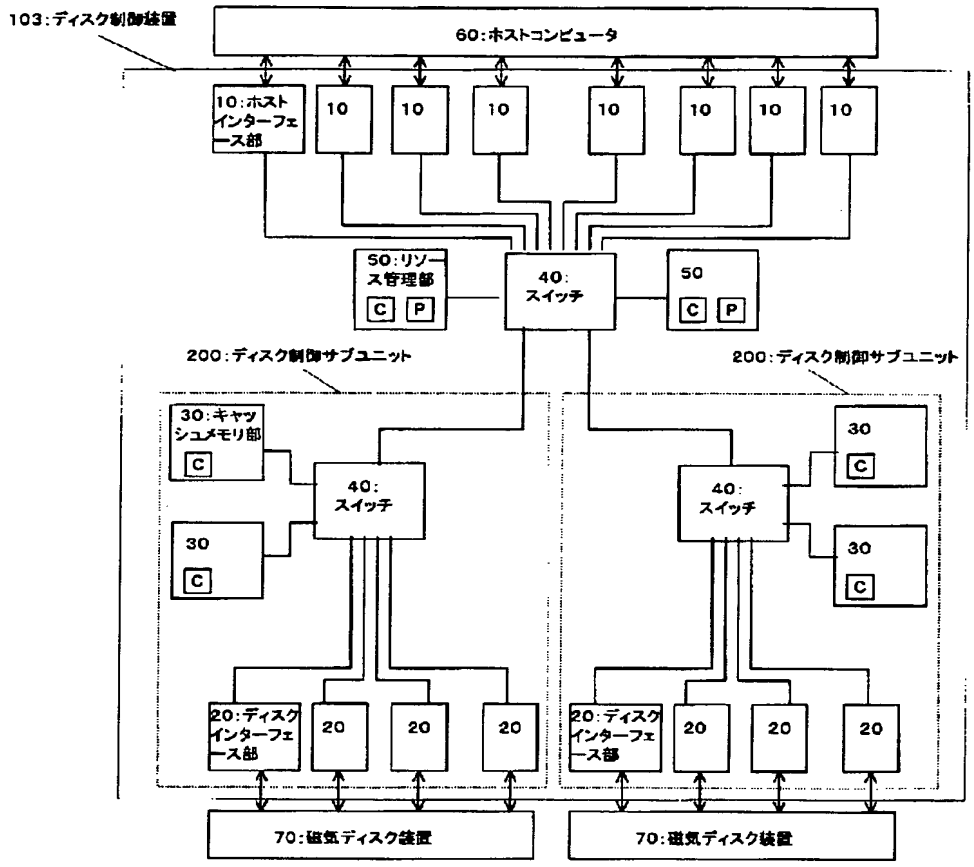
【図４】

図4


102:ディスク制御装置
60:ホストコンピュータ
10:ホストインターフェース部
10
10
10
40:スイッチ
40:スイッチ
50:リソース管理部
C
P
50:リソース管理部
C
P
30:キャッシュメモリ部
30:キャッシュメモリ部
40:スイッチ
40:スイッチ
20:ディスクインターフェース部
20
20
20
70:磁気ディスク装置

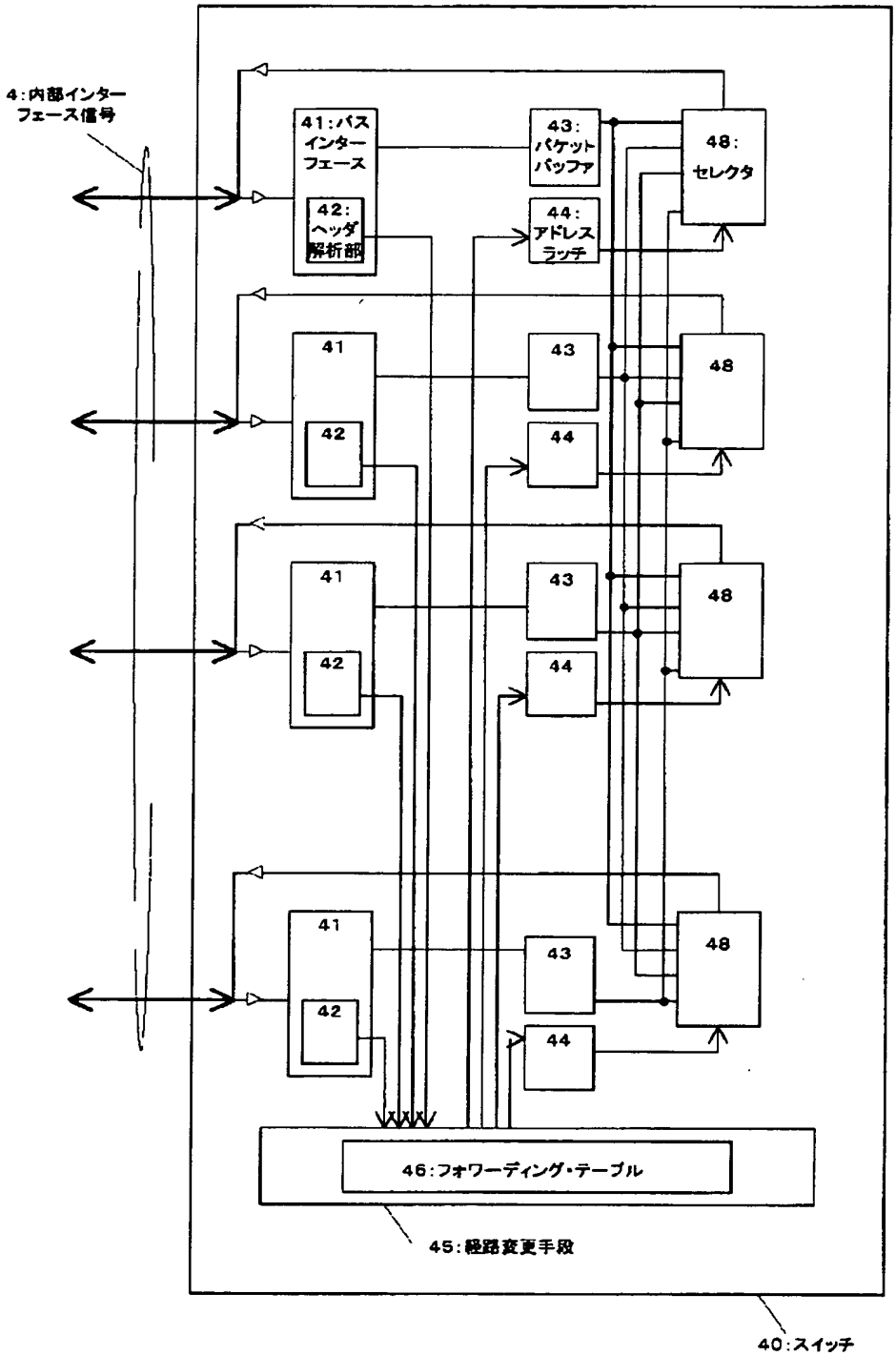
【図５】

図5


103:ディスク制御装置
60:ホストコンピュータ
10:ホストインターフェース部
10
10
10
10
10
10
10
50:リソース管理部
C
P
40:スイッチ
50
C
P
200:ディスク制御サブユニット
200:ディスク制御サブユニット
30:キャッシュメモリ部
C
40:スイッチ
30
C
40:スイッチ
30
C
30
C
20:ディスクインターフェース部
20
20
20
20:ディスクインターフェース部
20
20
20
70:磁気ディスク装置
70:磁気ディスク装置

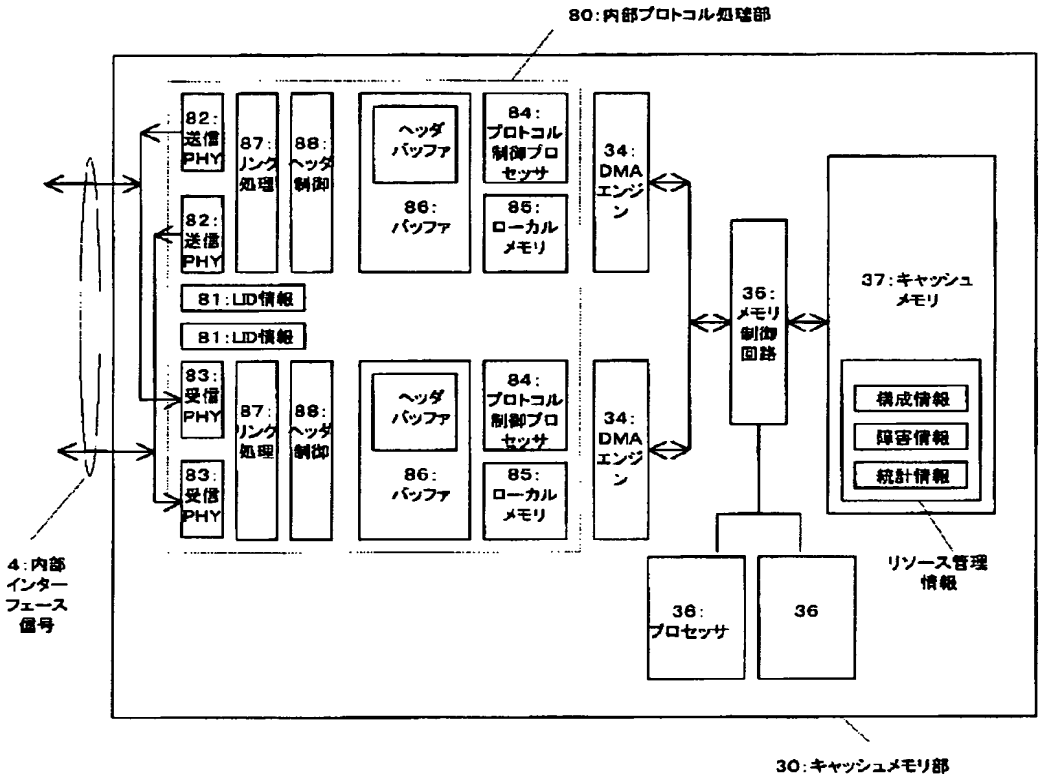
【図６】

図6


4:内部インターフェース信号
41:バスインターフェース
42:ヘッダ解析部
43:パケットバッファ
44:アドレスラッチ
48:セレクタ
41
42
43
44
48
41
42
43
44
48
41
42
43
44
48
46:フォワーディング・テーブル
45:経路変更手段
40:スイッチ

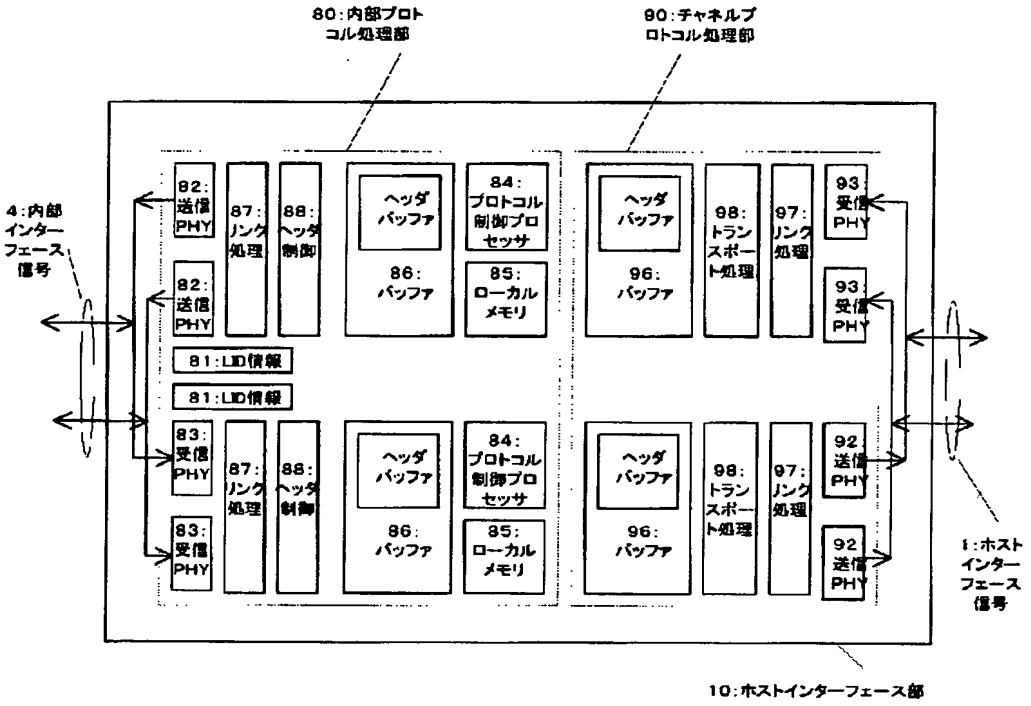
【図７】

図7


80:内部プロトコル処理部
82:送信PHY
82:送信PHY
87:リンク処理
88:ヘッダ制御
ヘッダバッファ
86:バッファ
84:プロトコル制御プロセッサ
85:ローカルメモリ
34:DMAエンジン
81:LID情報
81:LID情報
83:受信PHY
83:受信PHY
87:リンク処理
88:ヘッダ制御
ヘッダバッファ
86:バッファ
84:プロトコル制御プロセッサ
85:ローカルメモリ
34:DMAエンジン
35:メモリ制御回路
37:キャッシュメモリ
構成情報
障害情報
統計情報
4:内部インターフェース信号
リソース管理情報
36:プロセッサ
36
30:キャッシュメモリ部

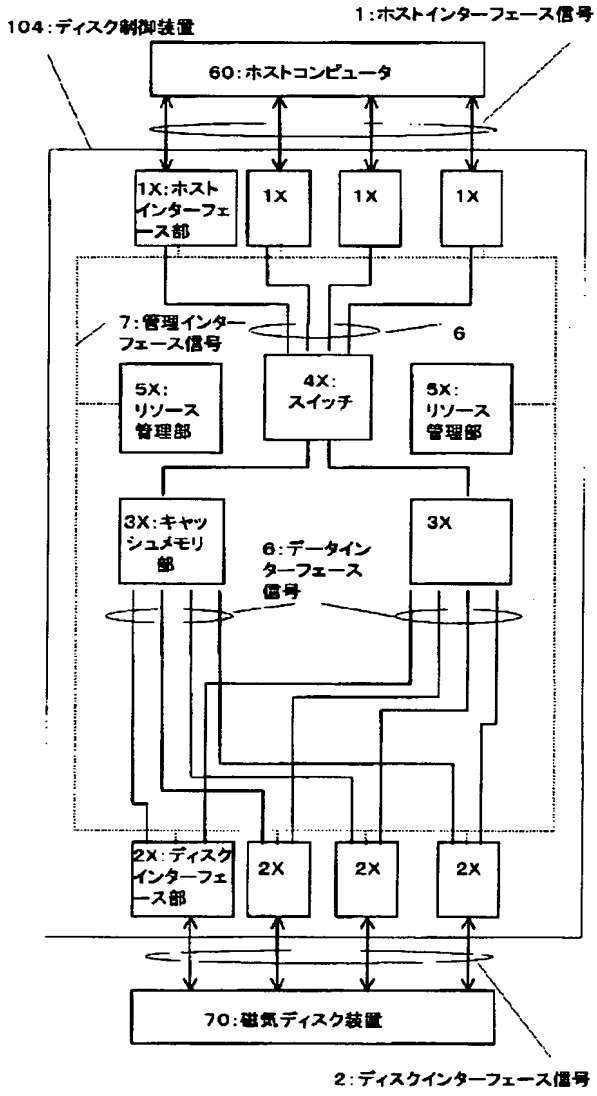
【図８】

図8


80:内部プロトコル処理部
90:チャネルプロトコル処理部
4:内部インターフェース信号
82:送信PHY
82:送信PHY
81:LID情報
81:LID情報
83:受信PHY
83:受信PHY
87:リンク処理
87:リンク処理
88:ヘッダ制御
88:ヘッダ制御
ヘッダバッファ
86:バッファ
ヘッダバッファ
86:バッファ
84:プロトコル制御プロセッサ
85:ローカルメモリ
84:プロトコル制御プロセッサ
85:ローカルメモリ
ヘッダバッファ
96:バッファ
ヘッダバッファ
96:バッファ
98:トランスポート処理
98:トランスポート処理
97:リンク処理
97:リンク処理
93:受信PHY
93:受信PHY
92:送信PHY
92:送信PHY
1:ホストインターフェース信号
10:ホストインターフェース部

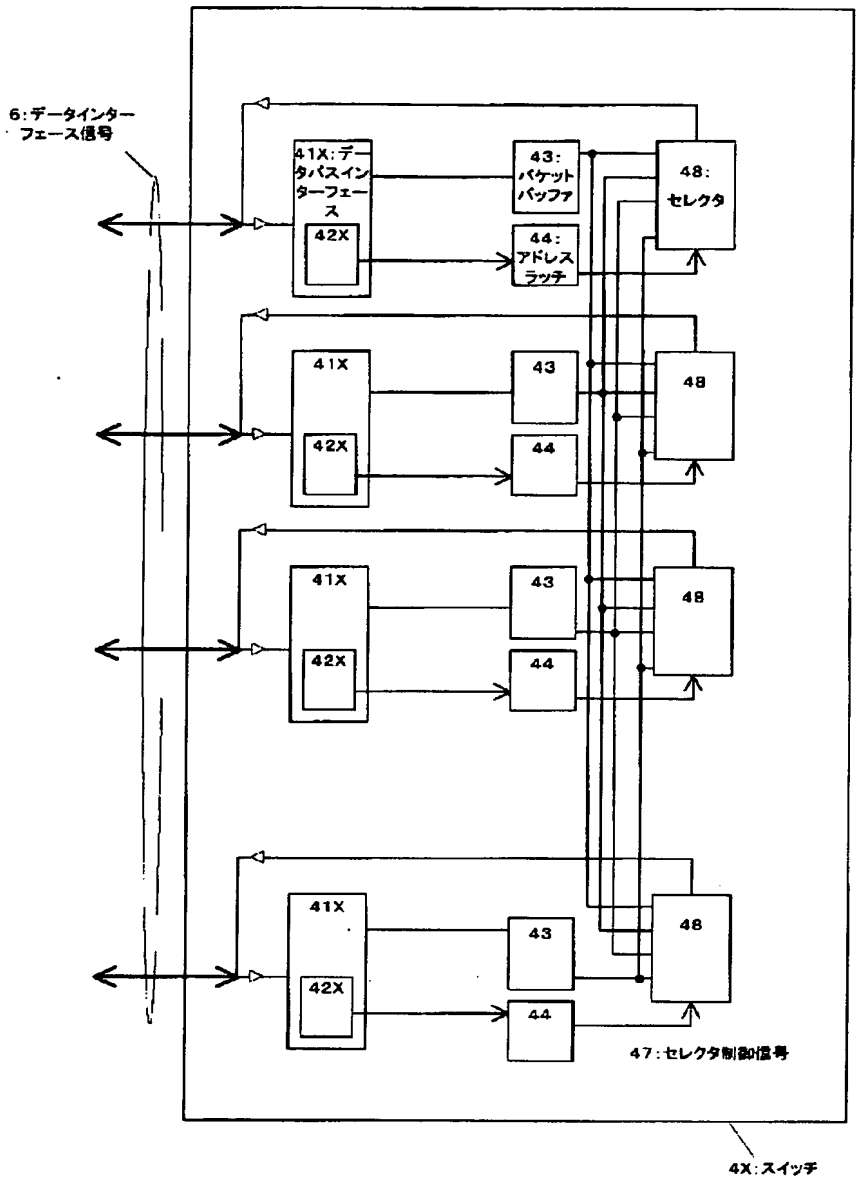
【図９】

図9


104:ディスク制御装置
1:ホストインターフェース信号
60:ホストコンピュータ
1X:ホストインターフェース部
1X
1X
1X
7:管理インターフェース信号
6
5X:リソース管理部
4X:スイッチ
5X:リソース管理部
3X:キャッシュメモリ部
6:データインターフェース信号
3X
2X:ディスクインターフェース部
2X
2X
2X
70:磁気ディスク装置
2:ディスクインターフェース信号

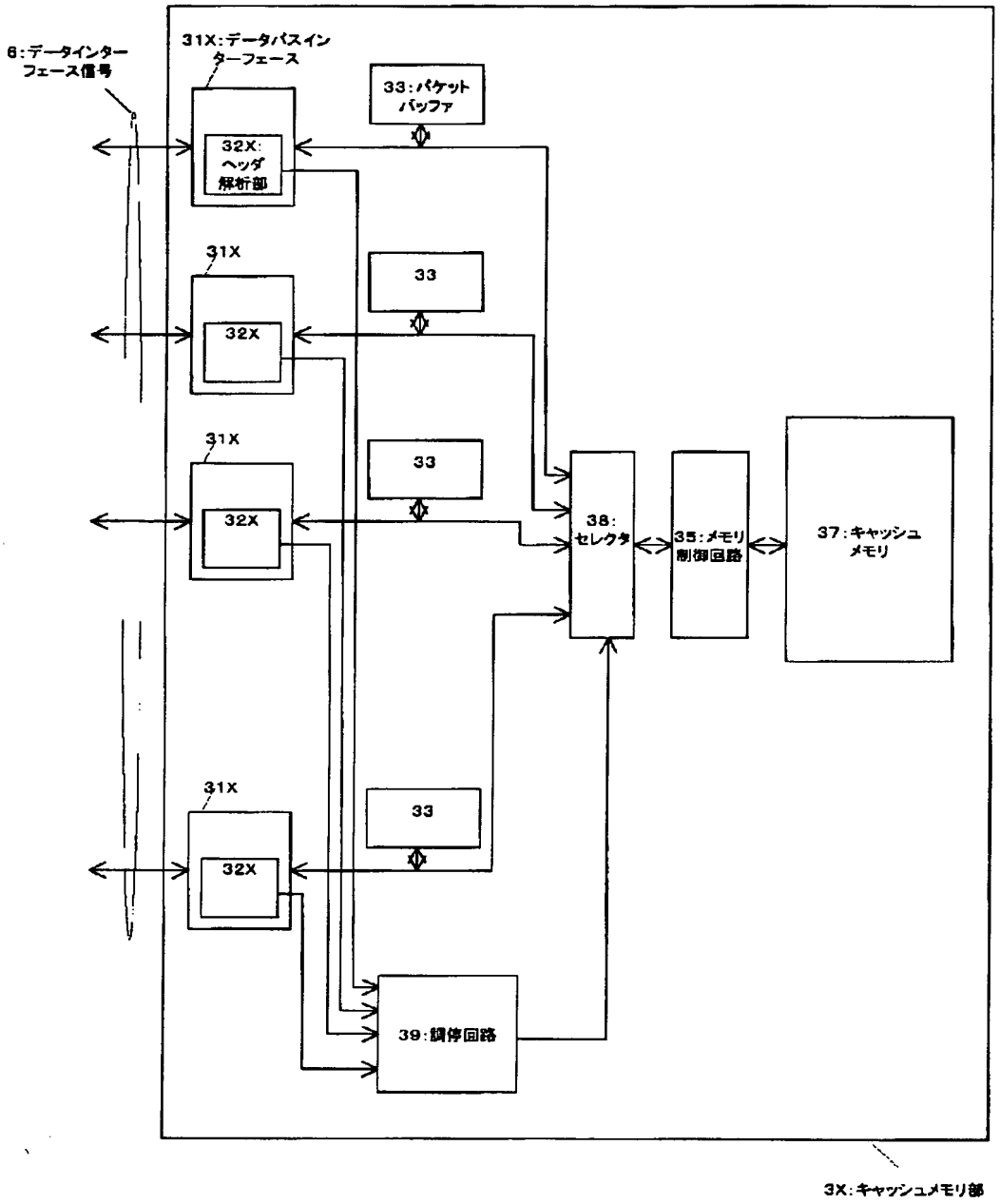
【図１０】

図10


6:データインターフェース信号
41X:データバスインターフェース
42X
43:パケットバッファ
44:アドレスラッチ
48:セレクタ
41X
42X
43
44
48
41X
42X
43
44
48
41X
42X
43
44
48
47:セレクタ制御信号
4X:スイッチ

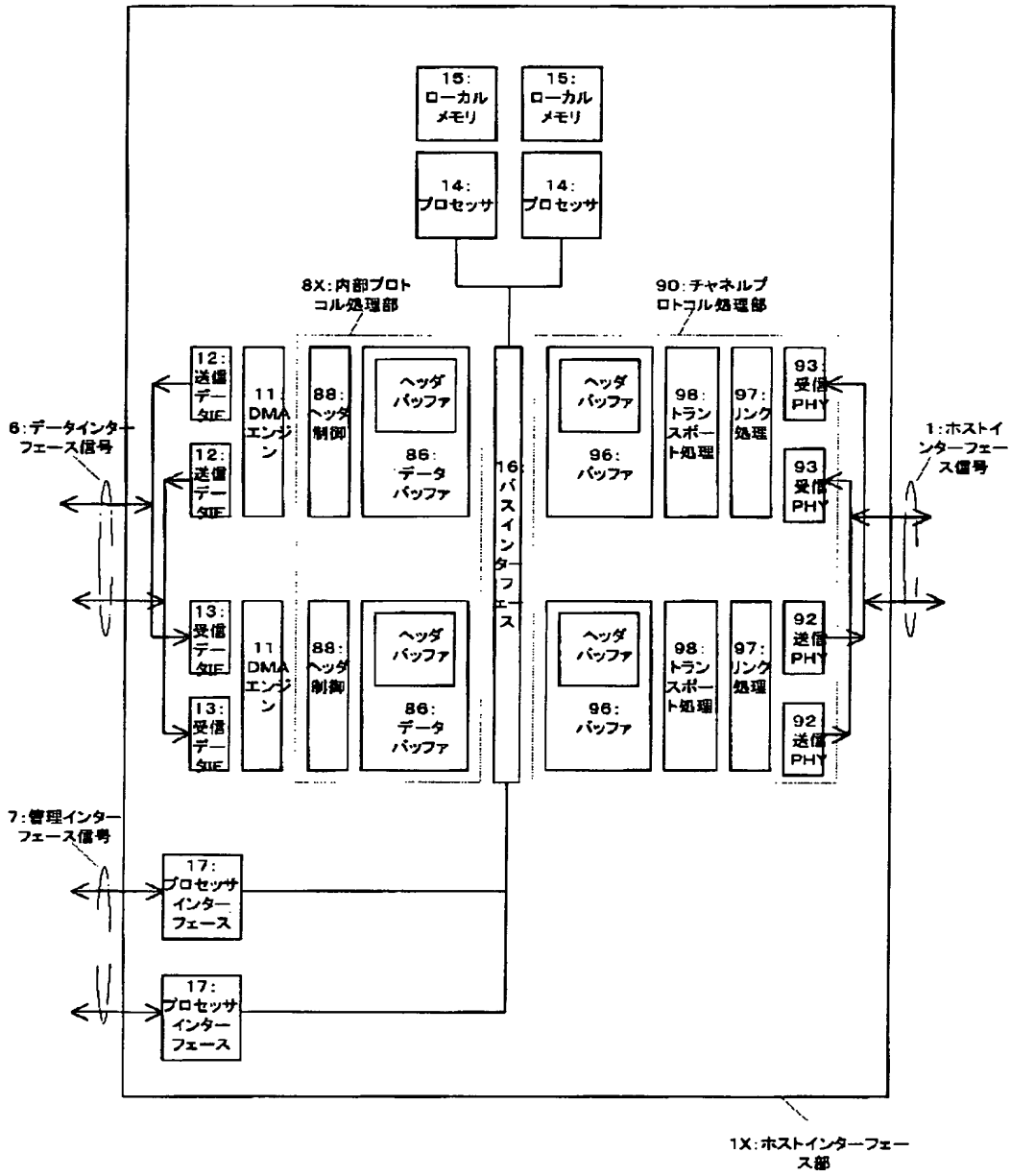
【図１１】

図11


6:データインターフェース信号
31X:データパスインターフェース
33:パケットバッファ
32X:ヘッダ解析部
31X
32X
33
31X
32X
33
31X
32X
33
38:セレクタ
35:メモリ制御回路
37:キャッシュメモリ
39:調停回路
3X:キャッシュメモリ部

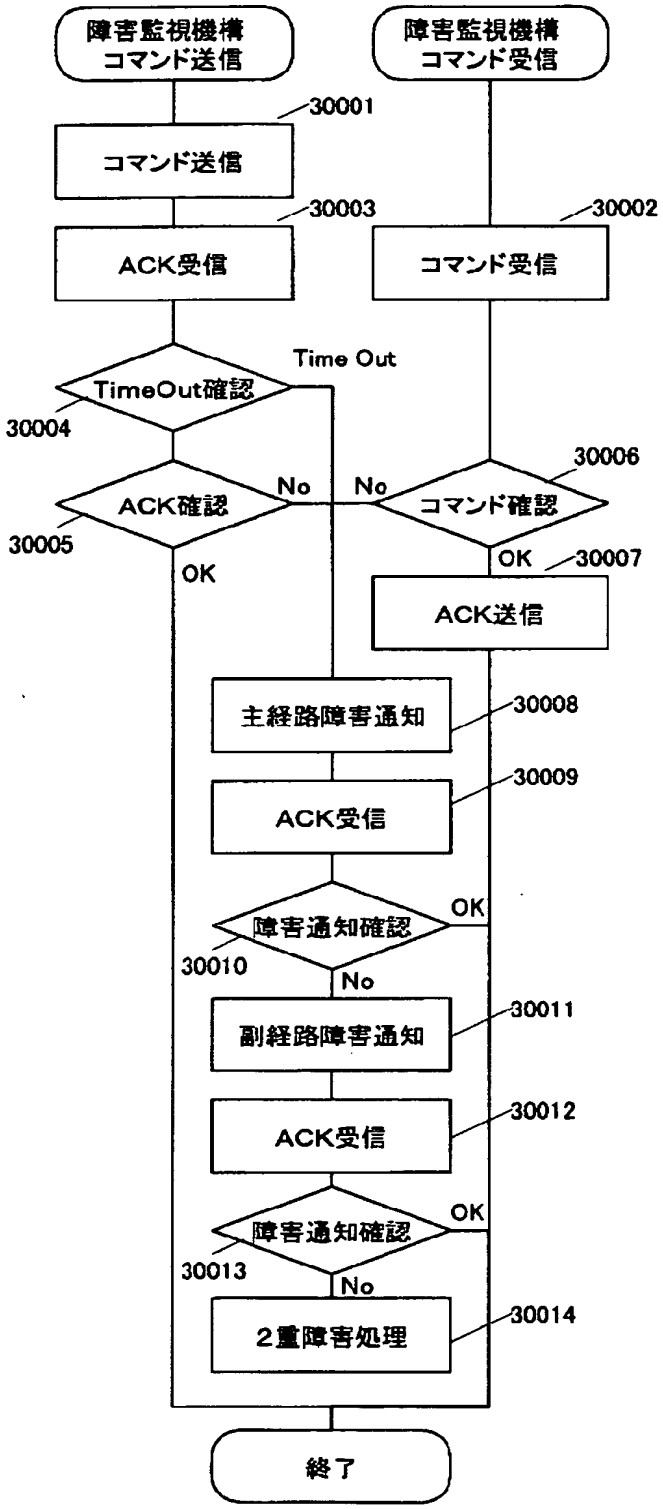
【図１２】

図12


15:
ローカル
メモリ
15:
ローカル
メモリ
14:
プロセッサ
14:
プロセッサ
8X:内部プロト
コル処理部
90:チャネルプ
ロトコル処理部
6:データインター
フェース信号
1:ホストイン
ターフェー
ス信号
12:
送信
データ
I/F
12:
送信
データ
I/F
11:
DMA
エンジン
88:
ヘッダ
制御
ヘッダ
バッファ
86:
データ
バッファ
16:
バスインター
フェース
ヘッダ
バッファ
96:
バッファ
98:
トラン
スポー
ト処理
97:
リンク
処理
93:
受信
PHY
93
受信
PHY
13:
受信
データ
I/F
13:
受信
データ
I/F
11
DMA
エンジン
88:
ヘッダ
制御
ヘッダ
バッファ
86:
データ
バッファ
ヘッダ
バッファ
96:
バッファ
98:
トラン
スポー
ト処理
97:
リンク
処理
92
送信
PHY
92
送信
PHY
7:管理インター
フェース信号
17:
プロセッサ
インター
フェース
17:
プロセッサ
インター
フェース
1X:ホストインターフェー
ス部

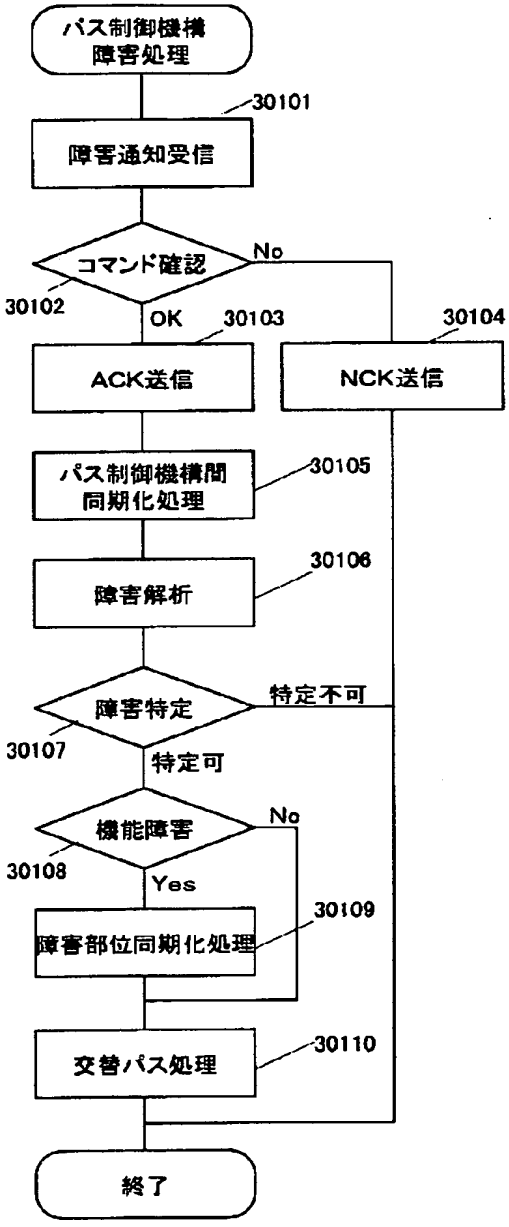
【図１３】

図13


障害監視機構
コマンド送信
障害監視機構
コマンド受信
30001
コマンド送信
30003
ACK受信
30002
コマンド受信
Time Out
TimeOut確認
30004
ACK確認
No
No
30006
コマンド確認
30005
OK
OK
30007
ACK送信
主経路障害通知
30008
ACK受信
30009
障害通知確認
OK
30010
No
副経路障害通知
30011
ACK受信
30012
障害通知確認
OK
30013
No
2重障害処理
30014
終了

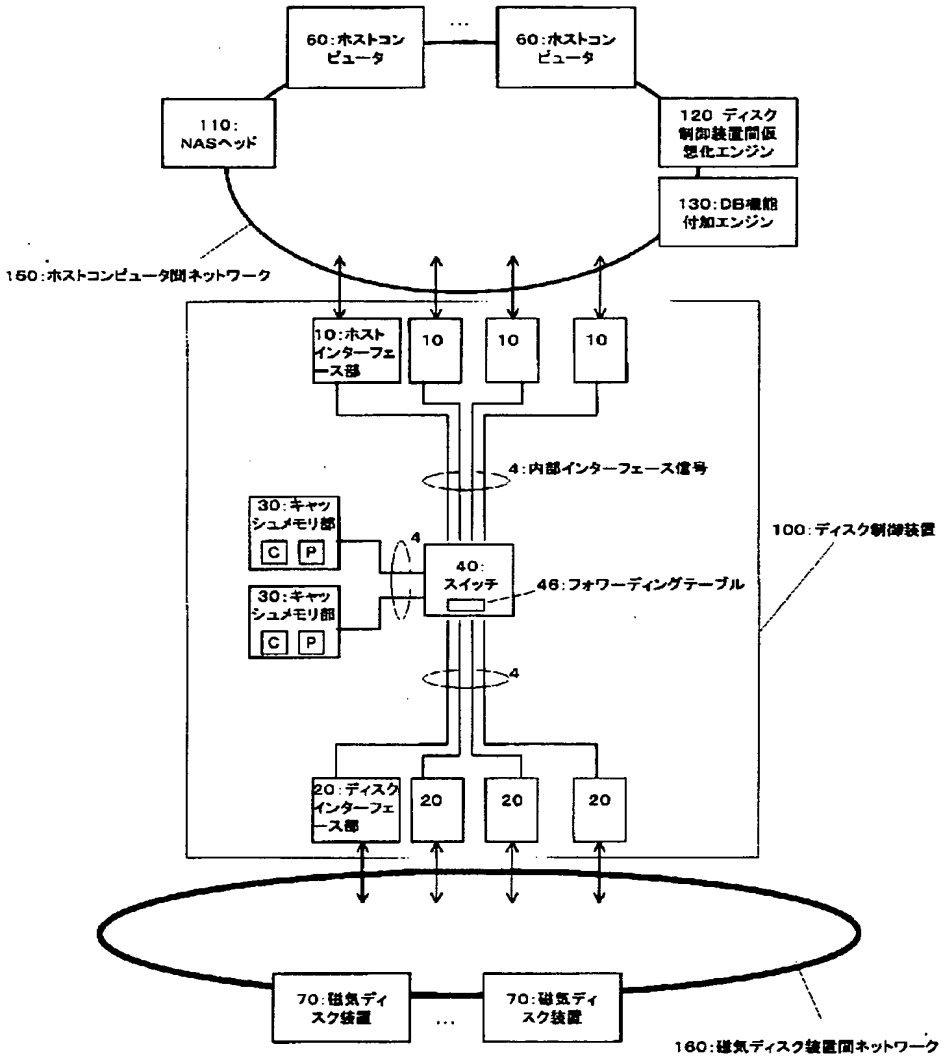
【図１４】

図14


パス制御機構
障害処理
30101
障害通知受信
コマンド確認
No
30102
OK
30103
30104
ACK送信
NCK送信
パス制御機構間
同期化処理
30105
障害解析
30106
障害特定
特定不可
30107
特定可
機能障害
No
30108
Yes
30109
障害部位同期化処理
交替パス処理
30110
終了

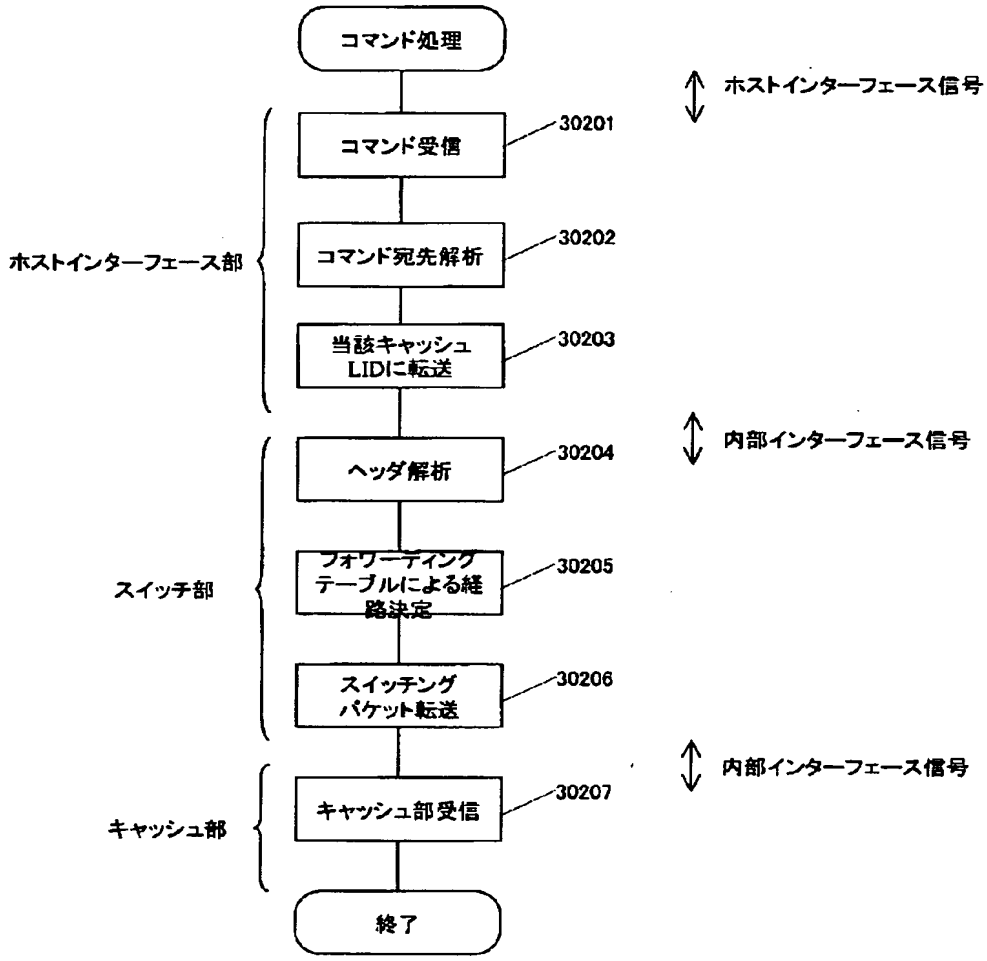
【図１５】

図15


60:ホストコンピュータ
60:ホストコンピュータ
110:NASヘッド
120 ディスク制御装置間仮想化エンジン
130:DB機能付加エンジン
150:ホストコンピュータ間ネットワーク
10:ホストインターフェース部
10
10
10
4:内部インターフェース信号
30:キャッシュメモリ部
C
P
4
40:スイッチ
46:フォワーディングテーブル
30:キャッシュメモリ部
C
P
100:ディスク制御装置
4
20:ディスクインターフェース部
20
20
20
70:磁気ディスク装置
70:磁気ディスク装置
160:磁気ディスク装置間ネットワーク

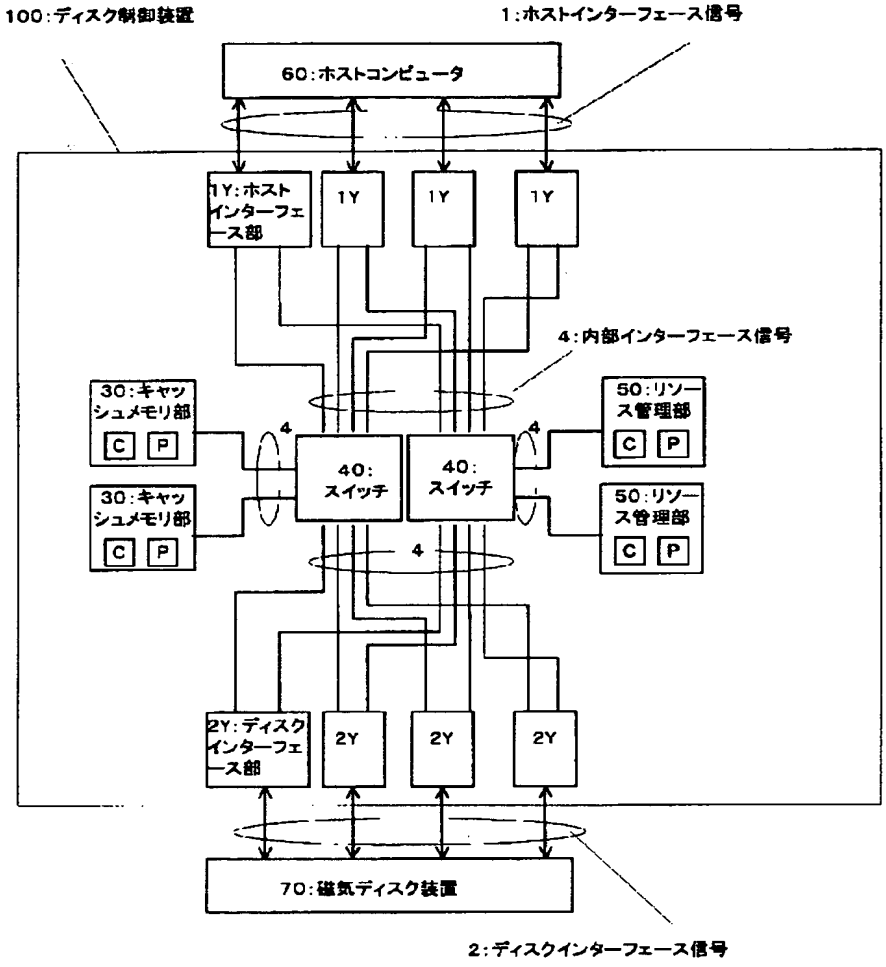
【図１６】

図16


コマンド処理
ホストインターフェース信号
コマンド受信
30201
コマンド宛先解析
30202
当該キャッシュLIDに転送
30203
ホストインターフェース部
内部インターフェース信号
ヘッダ解析
30204
フォワーディングテーブルによる経路決定
30205
スイッチング パケット転送
30206
スイッチ部
内部インターフェース信号
キャッシュ部受信
30207
キャッシュ部
終了

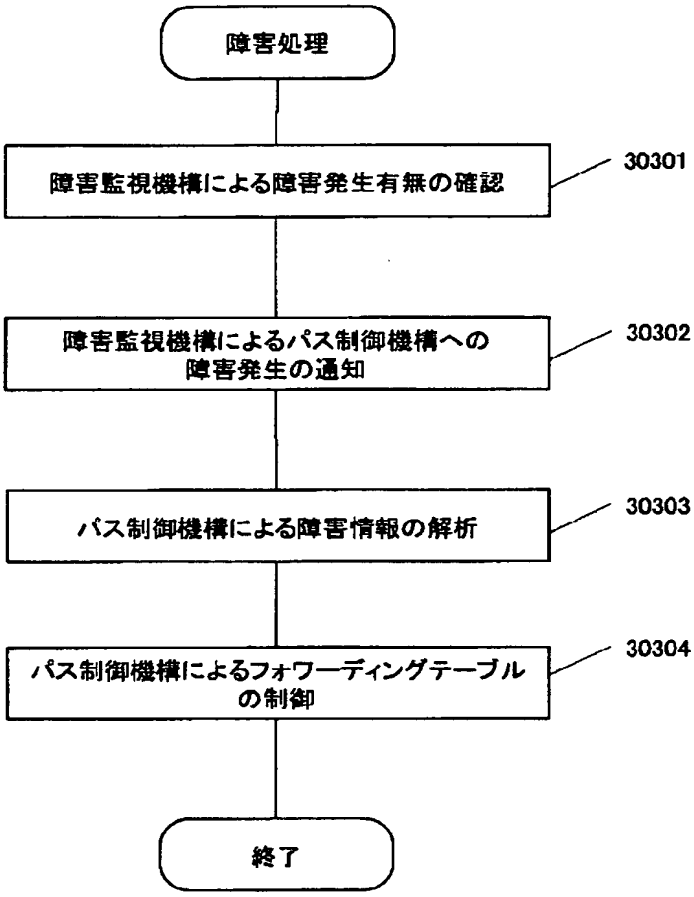
【図１７】

図17


100:ディスク制御装置
1:ホストインターフェース信号
60:ホストコンピュータ
1Y:ホストインターフェース部
1Y
1Y
1Y
4:内部インターフェース信号
30:キャッシュメモリ部
C
P
4
40:スイッチ
40:スイッチ
4
50:リソース管理部
C
P
30:キャッシュメモリ部
C
P
50:リソース管理部
C
P
4
2Y:ディスクインターフェース部
2Y
2Y
2Y
70:磁気ディスク装置
2:ディスクインターフェース信号

【図１８】

図18


障害処理
障害監視機構による障害発生有無の確認
30301
障害監視機構によるパス制御機構への
障害発生の通知
30302
パス制御機構による障害情報の解析
30303
パス制御機構によるフォワーディングテーブル
の制御
30304
終了

【書類名】　　要約書

【要約】

【課題】　ストレージ・システムの性能劣化や、ホスト・アプリケーションの動作不良を引き起こすことのない高可用性ディスク制御装置を提供する。

【解決手段】　複数のホストインターフェース部及び複数のディスクインターフェース部とキャッシュメモリ部との間を、１つ以上のスイッチで構成されるスイッチ網を介して接続し、スイッチにスイッチ網内での経路を指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を設け、また、ホストインターフェース部、ディスクインターフェース部及びキャッシュメモリ部に、スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を設け、さらに複数のキャッシュメモリ部に、その障害発生の有無を監視するための障害監視機構と障害発生時に障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を設ける。

【選択図】　図１

認定・付加情報

特許出願の番号	特願２００２－３７８９５６
受付番号	５０２０１９８１８０１
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１５年　１月　６日

＜認定情報・付加情報＞

【提出日】	平成14年12月27日

次頁無

出　願　人　履　歴　情　報

識別番号　　〔０００００５１０８〕

１．変更年月日　　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所